Prosodic Properties of Formality in Spoken Japanese

Ethan Sherr-Ziarko

St Cross College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Trinity, 2017

Dedicated to the memory of my grandmother, Claire Sherr, who supported me in so many ways throughout the process of completing this thesis.

Table of Contents

Acknowledgements

The list of people who must be thanked for helping me through the long (loooong) process of completing this thesis is extensive, but I'll do my level best to mention everyone here. First off, my family has been nothing but supportive in many ways of my endeavors at Oxford, and there is no way I would have completed my thesis successfully without them. Many, many thanks to my mom, dad, sister, uncle, and grandmother, all of whom kept me going in various ways. Just as important, without a doubt, was my supervisor John Coleman, who has guided me throughout my entire time in Oxford. Without his help my thesis would frankly have been a worthless load of rubbish, and I can't thank him enough.

Many people have helped with the gathering of the data presented in this thesis, but foremost among them and deserving of much gratitude is Kikuo Maekawa, who welcomed me to the National Institute of Japanese Language and Linguistics for a very productive research trip to Japan, and who also guided me through the process of creating a new corpus of spoken Japanese and has been helpful in myriad other ways. Thanks also to the many other kind and helpful people at NINJAL who made my field work a success.

I would not have remained sane throughout my time here without the many good times had with my friends at St Cross College. So, to Marten, Alex, Morgan, Kim, Jenny, Saher, David, Fraser, Hannah, Kimiko, and many others, thanks so much for all fun, drunken antics, gaming,

and bizarre conversations over the years. Life won't be the same without you. On the topic of St Cross, the amazing college kitchen staff has kept me fed and happy over the years, so thank you to all the folks there, past and present. Paul, Chris, Rob, Amanda, Yvonna, you are all fantastic and I probably would have starved without you.

Last but certainly not least, thanks to my wonderful girlfriend Nhung Nguyen, who has given me endless love, support, and inspiration (not to mention some damn tasty food) throughout all the ups and downs of a DPhil. I couldn't ask for a better partner in crime.

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

This thesis investigates the relationship between prosody and formality in spoken Japanese, from the standpoint of both language production and perception. This means that the primary research questions addressed by the studies described herein will not be concerned with how linguistic factors such as lexical items, morphological and syntactic structures, or social groups affect the realization of formality, but rather with the effects of prosodic variables such as fundamental frequency (hereafter f_0) and articulation rate. To frame this question in a real-world context, imagine a listener sitting in a noisy restaurant, listening to the conversation of two people sitting nearby. In such an environment, the listener would likely be unable to make out many of the words being used in the nearby conversation, but would perhaps be able to at least perceive some of the prosody of the conversation, such as the pitch contour and the speed at which they are talking. From just this information, would the listener be able to infer something about the level of formality of the overheard conversation? Conversely, from a production standpoint, would the speakers be able to convey something about their intended attitude and level of formality based not only on what they say, but on *how* they say it? Japanese is a particularly interesting language about which to ask these questions, as levels of formality in spoken Japanese are often clearly indexed by lexical items and syntactic structures (see Section 1.2.2 for in-depth discussion), allowing for the objective classification of utterances into different levels of formality, and

therefore presenting more straightforward conditions than most languages under which to investigate formality.

This introductory chapter will first discuss theoretical concepts which are core to the research questions and hypotheses being investigated in this thesis, including the properties of speech registers and styles in Japanese. This will be followed by a more in-depth discussion of the specific research questions and hypotheses regarding them that this thesis seeks to address. Finally, there will be an overview of the four content chapters to follow, which include: a pilot production study, a corpus-based production study, a speech perception study, and a description of a predictive statistical model of formality in spoken Japanese.

1.2 Speech register in Japanese

The first theoretical concept of importance for this project is that of *speech register.* Speech register here refers to sociolinguistic register, or more specifically to the broad level of politeness and formality that is being used in conversation (Halliday & Hasan, 1976). Different registers of speech are particularly well-defined in spoken Japanese, in that the register being used determines to a certain extent both a speaker's lemma selection in cases where variation is possible, as well as the use of syntactic and morphological forms (see Hinds, 1976). For example, a man using a formal register of speech would likely choose to use the 1st person pronoun /wataʃi/ ("I"), while

in a less formal situation he might choose to use /bokɯ/ (Sreetharan, 2004). A woman in the same situation, on the other hand, if using the "feminine" speech register, might instead choose to use the more stereotypically feminine 1st person pronoun /ataʃi/ ("I"), while in a less formal situation she might use /ɯtʃi/ (Siegal & Okamoto, 2003). An overview of previous work investigating Japanese speech register will give a better idea of its significance to the dynamics of the language.

1.2.1 Previous work

In general, previous linguistic research on the concept of speech register in Japanese can be broken down into two broad categories: studies looking at registers related to politeness and formality – which include registers such as *keigo* ("honorific language"), *kensongo* ("humble language"), semi-formal/polite speech, and informal speech, among others – and registers related to a speaker's social group – including registers such as gendered speech (Siegal & Okamoto, 2003; Inoue, 2002), businessmen's speech (Sreetharan, 2004), and dialects (Kubozono, 2012), among others.

Examinations of speech register in terms of politeness or formality in Japanese tend to be from the point of view of pragmatics (Pizziconi, 2002; Ide, 1982; Matsumoto, 1988); these studies often focus more on contrasting politeness strategies in Japanese with those presented in more universal frameworks of politeness (Brown & Levinson, 1987) than on the actual properties of different speech registers themselves. For the purposes of this study these pragmatic approaches

are of less interest than those that focus on the different acoustic and sociolinguistic aspects of politeness in Japanese, but before reviewing those sources there is one issue that is important to address, specifically the relationship between *politeness* and *formality* in Japanese.

1.2.1.1 Politeness versus formality in Japanese

Throughout this thesis, and occasionally in the previous literature (to be reviewed in Section 1.2.1.2), the concepts of *formality* and *politeness* are often conflated, or treated as roughly equivalent in Japanese. It is not immediately obvious that it is justifiable to do so, as there is nothing in particular stopping formal speech from being either polite or impolite in many languages (Brown & Levinson, 1978), and issues of formality are often ignored or minimized in general frameworks of politeness (such as Brown & Levinson, 1987). However, pragmatic studies of politeness in Japanese (Ide, 1982; 1989) have shown that there is a close relationship between formal speech and polite speech in Japanese.

Although Ide (1982) was primarily concerned with women's speech in Japanese, that paper also proposed a series of rules which related the polite and formal registers of spoken Japanese. Specifically, one overriding rule was proposed: that speakers should "Be polite in a formal setting" (Ide, 1982: 371). Notably, this rule applies only to *conversational* Japanese, and not strictly to written or scripted forms where informal and formal registers can sometimes combine in polite language. Ide (1982) proposes a relationship between formality and politeness wherein

formal speech is always polite (based on the overriding rule), but informal speech can be either polite or 'plain' depending on context and lemma selection. Examples of this contrast (after Ide, 1982) are given in (1).

(1) a) Suzuki-*si* wa kyonen Amerika e *irassyai masi-ta.* (formal, polite)
Mr. *last year* *to went*
(formal) (neutral) (formal)
'Mr. Suzuki went to America last year.'

b) Suzuki-*san* wa kyonen Amerika e irrasyai masi-ta *yo.* (informal, polite)
Mr. (*intensifier)*
(informal) (formal) (informal)

c) Suzuki wa kyonen Amerika e *it-ta* yo. (informal, plain)
went
(informal)

(1a) shows a typical formal, polite form of speech where formal forms of address ('si' rather than 'san' for 'Mr.') and verb forms are used. (2b) shows an example of how some properties of the informal register (such as the sentence final particle 'yo') can enter polite speech, while (1c) shows a sentence which is entirely in the informal register, drops the polite form of address ('san') and uses an informal form of the verb 'to go'.

The critical observation from Ide (1982) is that in Japanese formality appears to be a property of politeness, and the two registers are closely related on a pragmatic level. Similarly, while speech containing aspects of the informal register is not always considered impolite, speech which contains entirely informal lemmas and grammatical forms will rarely if ever be considered polite (Ide, 1989: 226-229). Because of this close relationship between polite and formal speech

in Japanese, previous work which focused primarily on politeness rather than formality is still highly relevant to the current study. Although the relationship between *informal* speech and politeness is slightly more nebulous, it is reasonable to expect similar results when studying the prosodic properties of formal speech as were seen when polite speech was investigated. Furthermore, based on the fact that informal speech seems to span different politeness categories, it is possible that analysis of the prosody of informal speech will produce a broad range of values that overlaps somewhat with the data observed for formal/polite speech.

1.2.1.2 Prosody of politeness/formality in Japanese

One classic study of the relationship between f_0 and politeness in Japanese speech is Loveday (1981). The study involved five Japanese speakers reading prompts while role-playing different situations, such as greeting a friend on the phone, speaking to clients, or speaking to their boss. The objective was to assess how speakers of different genders would use shifts in f_0 as a strategy to indicate politeness. Results indicated that female speakers of Japanese would use higher f_0 to indicate politeness, though such results did not appear clearly for men. Although this particular study is oft-cited as evidence that higher f_0 is a property of polite speech in Japanese, there are many issues with the study that make the results questionable. The largest such issue involves the measurement of f_0 itself – Loveday (1981) reported f_0 for males in the 65 – 400 Hz range, and for females in the 110 – 1050 Hz range, while the typical pitch range for male speakers

is ~50 – 200 Hz, and ~180 – 400 Hz for females (Hart et al, 2006). This indicates the presence of serious pitch estimation errors, which cast doubt upon any results of the study based on f_0. Secondly, no statistical tests were performed on the data, and all conclusions were drawn based on inference. Finally, the speech was all read and elicited via a task, meaning it is not an appropriate substitute for speech used in actual conversation, due to the differing prosodic properties of the two styles in Japanese (Nakamura et al, 2007). Therefore, although Loveday (1981) provides some possible evidence that higher f_0 may be a property of polite speech in Japanese, its conclusions are not convincing.

Another early study examining the relationship between speaking style and various prosodic parameters in speech collected in a laboratory was Sagisaka & Miyatake (1988). Although this study was not published in English, the results were presented in brief in the Journal of the ASA (1988); in the study, speakers (who were professional Japanese narrators) were prompted to speak written sentences in different 'styles' – such as fast, strong, high pitch, or low pitch – and were also asked to read additional sentences 'conversationally'. This study was conducted with an eye towards application of results on speech synthesis. The results indicated that f_0 and amplitude tended to be correlated, but that f_0 could also vary independently based on speech style. These results support the hypothesis that prosody and formality can co-vary in speech, but they should be approached with some caution. The analysis was entirely on read

speech collected in a laboratory, and previous studies have shown significant prosodic differences between read and spontaneous speech (Nakamura et al., 2007).

Another previous study of relevance is Ofuka & colleagues (2000), an acoustic experiment regarding the prosodic cues relating to different levels of politeness in Japanese. In this study, a series of (read) utterances were elicited from speakers in a controlled laboratory setting, and the subjects were asked to deliver the questions in both a 'formal' and then a 'casual' manner. Those recordings were then used in a perceptual experiment to test the effects of different acoustic cues on listeners' perception of an utterance's politeness. The study was one of the first to directly address the acoustic properties of casual versus formal speech in Japanese, and the target features examined were, perhaps necessarily, somewhat narrow – only the duration of the final vowel (used as a proxy for speech rate) and the properties of f_0 (in this case, mean, standard deviation, range, and rate of rise or fall in pitch) in the final mora.

The results of Ofuka et al. (2000: 213-215) from a perception standpoint were that – as foreshadowed by other studies of the perception of politeness in Japanese (e.g. Ogino & Hong, 1992) – both the direction of final f_0 movement and the speech rate of the final vowel were used by listeners as indicators of the intended politeness of an utterance. From a production standpoint the results were somewhat ambiguous – though the majority of speakers had a higher f_0 in the less polite utterances, one showed the opposite pattern. Speech rate, however, was consistently

higher in the less polite utterances. These findings are highly relevant to the current study, in that the possibility that f_0 and speech rate both co-vary with speech register is one of the hypotheses of this project (see section 1.3 for further discussion). However, Ofuka et al. (2000) does have some problems – firstly, only two sentences were tested, and both were questions, leading to a rather semantically narrow set of stimuli. Further, the fact that the recorded stimuli were all read rather than spontaneous speech, and were also collected in a laboratory is likely to have had undue influence on the results, in spite of all measures taken to avoid confounds. Therefore, although the study provides a solid theoretical basis to work from, the current study will attempt to overcome some of the problems with Ofuka et al (2000) by analyzing a broader range of more natural speech stimuli.

A third study, Ito (2002), the results of which contrast with Sagisaka and Miyatake (1988) and Ofuka et al. (2000), investigated the possible prosodic features of politeness in Japanese speech. It had the specific objective of addressing the problems with the unnaturalness of the speech used in the current studies on Japanese politeness and speech style (such as Sagisaka & Miyatake, 1988) and on examining the acoustic properties used by Japanese speakers to indicate a higher level of politeness. Although it was focused more on the properties of polite than informal speech, the objectives are still fairly similar to those of this study, and thus Ito (2002) can serve as a potential parallel.

Although Ito (2002) placed an emphasis on the naturalness of the speech data used, the data was still elicited via a task – wherein speakers gave each other directions to locations on a map – in a laboratory environment, and so the naturalness of the speech is somewhat debatable. Still, it likely came closer to being spontaneous than the data from studies such as Sagisaka and Miyatake (1988). Although it used only four speakers, interestingly the different speakers appeared to have different strategies for indicating formality; one speaker raised f_0 to indicate greater levels of politeness, while another showed no such change (Ito, 2002: 2). Also, in a perception experiment, listeners did not appear to use changes in f_0 to assess increases in formality. These results emphasize the potential importance of individual speaker style to the properties of different speech registers that this study is focused on, but also shows some potential problems relevant to the current study. It was clear from the results of Ito (2002) that listeners place a high value on semantic and morphological cues to indicate levels of formality (as they should, given how indexical many lemmas/morphemes in Japanese are), and so care will have to be taken when designing any stimuli for perception experiments.

It is worth noting however that Ito's (2002) results are not directly applicable to this study; Ito was focused strictly on formal speech, ranging from a polite register, to full honorific (*keigo*). The results showing only weak acoustic indicators of formality are therefore – while still worth noting – of only limited applicability to this study, given the differing focuses.

1.2.1.3 Prosody of formality in other languages

There are a number of studies of the relationship between prosody and formality in languages other than Japanese (such as Spanish, German, Korean, and English; see Navarro & Nebot, 2014 for an overview). Of these, the study which is most likely to be relevant to the current study is Winter & Grawunder (2012) which investigated the relationship between a number of acoustic factors and formality in Korean. The reasons for its particular relevance are that, firstly, it is concerned with *formality* rather than *politeness,* which mirrors the current study. As the relationship between formality and politeness in Korean is not necessarily as clear as it is in Japanese, this is a critical point. Secondly, it has been observed that there are some similarities in prosodic structure between Korean and Japanese, particularly at the level of the intonational phrase (Kubozono, 2015), and so making comparisons between the prosody of the two languages is not entirely far-fetched.

Winter & Grawunder (2012) collected speech of different levels of formality via a role-playing task, where subjects were asked to either leave a message on a cell phone voice mail, or to make a direct request of someone in person. Both scenarios were used to produce examples of formal and informal speech. The acoustic properties – including mean, range, and SD of f_0 and intensity, harmonics, articulation rate, pause count, filler count, and breath intakes – of the different levels of formality were analyzed and compared using mixed effects regression models.

Significant main effects were found for mean, SD, and range of f_0 ($p < .01$ for all), articulation rate ($p < .05$), and filler count ($p < .001$), wherein all were significantly higher in informal than in formal speech.

Although the results of Winter & Grawunder (2012) are not necessarily enough by themselves to hypothesize that the same results will be seen in the current study (as Japanese and Korean are not related languages), the observed similarity between the prosody of the two languages allows them to serve as a point of reference. The fact that many aspects of f_0 (mean, SD, and range) appear to co-vary significantly with changes in formality indicates that it may be worth investigating the relationship between f_0 and formality in Japanese to a similar (or greater) level of depth. Additionally, the fact that many of the tested prosodic variables were higher in informal speech is of interest, as this result can help better inform the hypotheses of the current study as they relate to the expected relationship between prosody and formality (see Section 1.3.2 for further discussion).

1.2.1.4 Social group

The aspect of speech register in Japanese relating to social group that has seen by far the most study is the issue of gendered speech. The primary reason for this preponderance of study in a particular area is likely that Japanese is, while not entirely unique, still fairly singular in that it contains numerous words and morphological units which are exclusively indexical of gender,

for both “masculine” and “feminine” speech (Ide, 1982; Sreetharan, 2004; Inoue, 2002; Siegal & Okamoto 2003). As this thesis is not concerned specifically with issues of speech register related to social group or gender I will not give an in-depth discussion of such registers (for discussions of gendered speech in Japanese, see e.g. Ide 1982; Inoue, 2002; Ohara, 2004; Seigal and Okamoto, 2003; for social group see e.g. Loveday, 1986; Sreetharan 2004; Kubozono 2012). However, there are some aspects of the prosody of these registers which is relevant to the design of some of the experiments in this thesis.

Although most work on gendered language in Japanese has been focused on its sociolinguistic or structural aspects, a small amount of work has been done regarding the prosody of women's language, in particular Hiramoto (2010). The main finding of Hiramoto (2010) of relevance to this thesis is that while Japanese women's speech involved speakers using a higher pitch (f_0) on average than male counterparts, f_0 was raised even more significantly after 'feminine' sentence-final particles such as 'ne', and less so after more 'masculine' sentence-final particles such as 'yo' (Hiramoto, 2010: 113-116). This indicates that there are effected changes in prosody used by speakers that are indexical of a feminine register of speech. For this reason, it will be particularly important to take account of gender differences in any analyses of f_0.

One further study of gendered speech that is important to the current study is Ohara (2004), which examined differences in f_0 range between men and women when producing speech in the

same social situation (such as asking someone to wait). Although no statistical analyses were performed, Ohara (2004) found that female speakers used a much wider f_0 range than males based on observation of pitch contours and of the difference between the min, max and SD of f_0 in a number of different utterances. Although the analysis of f_0 range was not particularly rigorous, the findings in Ohara (2004) do demonstrate that it will be important to take speaker gender into account in any potential investigations of f_0 range in the current study, in addition to the more well-attested gender based differences in mean f_0.

A final area of study regarding Japanese speech registers which is worth noting is the study of the numerous Japanese dialects. Essentially every region of Japan has its own dialect, some of which are mutually incomprehensible (Kubozono, 2012), and which often bring with them large-scale changes to tonal patterns, lemma selection, and syntax (Kubozono, 2012). While the existence of these dialects must be acknowledged, as they are a large part of daily spoken Japanese and therefore could be considered as a type of informal speech register, the sheer number of confounds considering them in this study would present led to the decision to exclude the examination of dialect differences entirely. In spite of the fact that a similar study concerning dialects would undoubtedly be interesting, it would unfortunately be outside of the scale of this project – only 'standard' Japanese as spoken by natives of the Tokyo region will be considered.

1.2.2 General overview of the properties of Japanese speech registers

For the purposes of this thesis, it will be useful to have a brief general summary of some of the syntactic, semantic, and lexical properties of different registers of speech.

1.2.2.1 Formal speech

Formal but non-honorific (see Hori, 1986) speech in Japanese is characterized by the selection of certain lemmas, morphological units, and syntactic structures. Perhaps the most consistent and immediately obvious property of more formal speech is the common verb ending /-masɯ/ (Cook, 1998). In morphological terms, this ending will simply attach itself to any verb root (such as /tabe-/ "to eat") to realize the "Complete" form of the verb. It is worth noting that this particular phenomenon is almost entirely related to speech register – there is no verb agreement in Japanese, and the different verb endings marking tense, voice, etc. (which will be discussed further in the section on casual speech) are independent of the subject or objects of a given sentence (see Hinds, 1978 for an overview of verb forms in discourse). Use of the /-masɯ/ (or the negative /-masen/ or the past-tense /-maʃita/) verb ending is entirely determined by speech register.

Similarly, a formal register of speech encourages the selection of certain more formal lemmas – or, perhaps more accurately, disallows the use of more colloquial lemmas, as it is still generally permissible to occasionally select more formal lemmas while using an informal register

than it is to use informal lemmas in a formal register (Ide, 1982). For example, in a formal register of speech, a speaker would be much more likely – perhaps even required, depending upon the social situation at hand – to select the more polite verb meaning "to eat": /tabeɽɯ/ rather than the much more informal /kɯ:/. Additionally, Ide (1982) proposes that a feature of the formal register is the more frequent selection of Sino-Japanese words (those words borrowed from Chinese) rather than their *yamato* (native Japanese) homonyms (e.g. the selection of /ɽju:tʃo:/ 'fluent' vs the selection of /peɽapeɽa/ 'fluent'). There are many hundreds of such examples of lemma pairs which are more or less likely to be selected based upon the register being used (Table 1.1 shows examples).

There are also syntactic markers of the formal speech register in addition to those relating to morphology and semantics. Likely the most notable syntactic property of a more formal register of speech is a more constrained word order as compared to more informal speech (examples (2) and (3) below illustrate an example of this). This has been attested by probabilistic studies of Japanese word order (Hawkins, 2006), although it is not as completely consistent as the previously mentioned properties, as word order is somewhat (though not entirely) flexible in Japanese anyway. As Japanese does have a case system, it is generally permissible to change the order of phrases, or to place a phrase at the beginning of a sentence using the topicalizer /wa/. However, it does generally still hold that the verb must come at the very end of the sentence (or at least at

the end of the V'), whereas in more informal speech subjects or objects are occasionally moved to the end of the sentence. For example, in standard, formal Japanese it would be common to say

(2) /soɽe + o mimaʃita/

that + ACC see-PAST "I saw that"

whereas in colloquial speech it would be permissible (although perhaps beyond the limits of prescriptive Japanese grammar) to say

(3) /mita jo soɽe/

see-PAST (Intensifier) that (also "I saw that")

The slightly stricter word order in (2) is typical of a formal speech register in Japanese, while greater flexibility is conversely a property of an informal register (to be discussed further in this section).

As can be seen in examples (2) and (3), another point of difference between formal and informal registers is that more formal speech tends to either lack sentence final particles (such as the intensifier /jo/ in 2), or use a much more restricted set of SFPs (see Ide, 1982; Sreetharan, 2004 for a more in-depth discussion of SFPs as related to speech register). In general speech in the formal register will be limited to using SFPs with some grammatical or semantic function, such as, for example the question particle /-ka/ or the quotative particle /-to/, rather than those that carry some paralinguistic or pragmatic information. It is worth noting that it is still entirely

possible to have a number of different SFPs in *polite* speech (Ide, 1982: 372), and it is one of the few areas in which formal and polite speech registers diverge in Japanese.

Table 1.1: *Properties of typical formal speech in Japanese.*

	Phenomenon	**Example**	**Example**
Syntactic Structures	Word Order	/ʃaʃ:in + o toɾimaʃita/ picture + OBJ take-PAST "I took a picture"	/tot:a jo ʃaʃ:in/ take-PAST (intensifier) picture "I took a picture" (informal)
Lemmas	Polite lemma selection	/taberɯ/ "to eat" (polite) vs. /kɯ:/ (informal)	/wataʃi/ "I" (polite) vs. /oɾe/ (informal)
	Sino-Japanese lemma selection	/ɽju:tʃo:/ 'fluent' (formal) vs /peɾapeɾa/ (informal)	/ho:be:/ "visit America" (formal) vs. /ameɾika + e ikɯ/ (informal)
Verb endings	Verb endings	/kakimasɯ/ "to write" (formal) vs. /kakɯ/ (informal)	/mimasen/ "see-NEG" (formal) vs /minai/ "see-NEG" (informal)
Sentence final particles	Lack of SFPs	/kaɾe + ga ameɾika + e ikimaʃita/ he + SUBJ America to go-PAST "He went to America" (formal)	kaɾe + ga ameɾika + e ikimaʃita + jo/ he + SUBJ America to go-PAST (Intensifier) "He went to America" (informal)

1.2.2.2 Informal speech

Finally, the properties of the informal register of speech must be examined, as they are of the greatest relevance to this project. As previously discussed, the exact properties of the informal speech register in Japanese are somewhat nebulous (as they are in most languages), and are

somewhat dependent on individual speaker style. However, there are a number of concrete properties that are indexical of informal speech in Japanese. In particular, there is a marked tendency to drop or shorten various morphological or phonological features during casual speech – for example, it is quite common for case particles to be dropped during informal speech, even as word order becomes slightly freer as was previously discussed. For example:

(4) a) *ore-wa ima ikimasu*

I-TOPIC now go "I'm going now." (polite)

b) *ore ima iku*

I now go "I'm going now." (informal)

In (4), the first, less casual utterance (4a) retains the topicalizing particle /wa/ after the subject (*ore* "I"), while in the second, more informal sentence (4b) this particle is entirely absent. As with many properties of the informal register this omission of morphological items is not universally permissible or always done, but such alternations are virtually non-existent in more formal speech, and so this omission of particles by speakers can be used to categorize utterances as informal.

As can also be seen in (4), another property of the informal register is the frequent use of the infinitive of verbs, without the *-masu* verb ending. The past tense and negative forms of verbs are also different in less formal speech – examples of which were given in Table 1.1. There is

also a very broad difference in the selection of lemmas between the polite and informal registers, examples of which were given in Table 1.1. Another property of informal Japanese speech which is of particular note to this thesis and will therefore be discussed in greater depth in Section 1.2.2.3 are the phenomena of elision and geminate contractions.

As discussed in Section 1.2.2.1, an informal register of speech also allows for the use of a much broader range of sentence final particles (Ide, 1982). Specifically, it is common in the informal register of Japanese for speakers to make frequent use of particles which contain paralinguistic or pragmatic information, such as intensifiers (e.g. /jo/, /ze/, /ɽo/, /na/), or interrogatives (e.g. /ne/). Again, these SFPs do not equate to speech being impolite, but are rather a property of the formal/informal categories.

1.2.2.3 Geminate Contractions

Elision and lengthening phenomena in spontaneous, informal Japanese speech have not been widely studied in the literature, but they have been observed appearing frequently in corpora of spontaneous Japanese (Arai, 1999; Kawahara, 2015). These phenomena will be discussed in greater detail in Chapter 2, where they are part of the experimental design, but in short speakers tend to elide vowels in a number of phonological conditions in spoken Japanese (Kawahara, 2015), which is typically then followed by a form of compensatory lengthening where the consonant following the elided vowel appears to geminate (Arai, 1999), although there is

shortening of the overall duration of the word.

Thereby, geminate contraction refers to the phenomenon in Japanese wherein speakers employ geminates post-lexically in words where they would not otherwise occur, occasionally violating canonical phonological principles of Japanese (Mester & Ito, 1995) such as the prohibition against voiced geminates in *yamato* (non-borrowed native Japanese) words, as in the example [wakan:ai] seen in Table 1.2. For example, it is very common – almost ubiquitous – in spoken Japanese for the word /atatakai/ "Warm" to be realized as [at:akai]. Table 1.2 gives a non-exhaustive list of examples of this phenomenon in Japanese.

Table 1.2: *Examples of geminate contractions.*

Standard Form	**Stylistic Geminate**
/atatakai/ "warm"	[at:akai]
/dokoka/ "somewhere"	[dok:a]
/wakaɾanai/ "don't understand"	[wakan:ai]
/tsɯmaɾanai/ "boring"	[tsɯman:ai]
/ɯɾɯsai/ "shut up"	[ɯs:e:]
/oɾe no ɯtʃi/ "my house/family"	/oɾentʃ:i/

The list in Table 1.2 is not nearly complete, as this a somewhat productive phenomenon, but it does give a few of the more common examples. It is difficult to draw any clear generalizations about geminate contractions; they do generally occur at morpheme boundaries, but not always (word internally in /atatakai/). They do not occur exclusively based on the presence

of repeating consonants in an auto-segmental tier – /oɽe no ɯtʃi/ → [oɽentʃːi] seems to involve only vowels eliding – and it is also not restricted to any particular part of speech. Perhaps the only generalization that can be drawn is that it only occurs in words with a very high lemma frequency; all of the examples in Table 1.2 are words with a lemma frequency in Japanese of at least 100 parts per million (according to the BCCWJ word frequency list; Maekawa et al., 2014), which while not incredibly frequent would nonetheless be among some of the more commonly used words in the language. Furthermore, in spite of the standard linguistic description of a geminate as a "long consonant" (Scheine & Steriade, 1986), geminate contractions appear to be used, almost exclusively, as a shortening tactic to reduce articulatory complexity. This observation comes from the fact that the phenomenon of geminate contraction tends to involve the elision of full CV moras.

Geminate contractions are important to this project primarily because they are one of the more direct indicators that an informal register of speech is being used. As will be discussed further in Chapter 2, the presence of stylistic geminates is a major criterion for the inclusion of recordings in the first experiment described in this thesis.

1.3 Research questions and hypotheses

1.3.1 Research questions

As was stated in Section 1.1, the primary research question of this thesis is *how is prosody related to the production and perception of formality in spoken Japanese*? This broad question leads to further, more specific lines of inquiry. The initial question to be addressed in the thesis approaches the issue from a production standpoint: (1) *Do speakers of Japanese make use of changes in prosody to help communicate their intended level of formality in conversation?* And if that is so, it leads to the further question (2) *what* specific *differences in the prosody of an utterance are used by speakers to distinguish different levels of formality?* Addressing these two questions necessarily precedes any investigation of questions involving speech perception, as without an idea of the prosodic variables involved in the expression of formality and the degree to which those variables are related to formality, it would be very difficult to design an appropriate experiment to test the salience of prosody to the perception of formality. Once these questions have been investigated, however, it will be possible to move on to a further research question: (3) *do listeners make use of certain prosodic cues to help determine the intended level of formality of the speaker?* This question can be investigated using information gained from experiments designed to address the first two research questions. Once the relevant prosodic variables are determined, it will be possible to manipulate them in order to test the relationship between the

prosody of an utterance and how its level of formality is perceived (see Section 1.4.5 for further details). Finally, after addressing these three questions, the final question that this thesis will investigate is: (4) *given the results of the experiments investigating questions (1)-(3), is it possible to build an accurate predictive statistical model of formality in Japanese?* This question will not be addressed experimentally, but rather in a Bayesian statistical framework which will be described further in Chapter 5.

1.3.2 Hypotheses

For each of the research questions posed in Section 1.3.1, experiments will be designed to test the hypotheses:

(1) *Speakers do use changes in prosody to express different levels of formality in speech.*

This hypothesis is based on both previous studies of Japanese (Loveday, 1981; Sagisaka & Miyatake, 1988; Ofuka et al., 2000; Ito, 2002) which showed significant relationships between prosody and the expression of formality, as well as on studies of other languages that showed similarly significant relationships (see Navarro & Nebot, 2014 for an overview). Although the results of these studies have not always been clear, or consistent even within individual languages, the literature does appear to support the hypothesis that prosody and formality are related on the level of speech production. In contrast to previous work however, the current study will test this hypothesis using a corpus of natural speech, so it is possible that it will produce very different

results.

(2) *Differences in f_0, articulation rate, and f_0 range will co-vary significantly with the level of formality in speech. Each of these variables will be higher in informal speech.*

This hypothesis is based upon the results of previous studies (Ofuka et al, 2000; Sagisaka & Miyatake, 1988) which showed significantly higher f_0 (Sagisaka & Miyatake, 1988) and speech rate (Ofuka et al., 2000) in 'casual' speech. Furthermore, a study of the relationship between prosody and formality in Korean (a language which has been described as having prosodic structural similarities to Japanese; see Kubozono, 2015) also showed a pattern where mean f_0, articulation rate, and f_0 range were all higher in informal than formal speech (Winter & Grawunder, 2012). Although this hypothesis contradicts the idea (based on studies such as Loveday, 1981; Ohara, 2001) that f_0 is higher in polite speech in Japanese, the lack of experimental and statistical rigor behind that idea combined with the results of previous studies showing increased f_0 and speech rate in informal speech makes the hypothesis justifiable.

(3) *Changes in the prosody of an utterance will have a relationship to listeners' category judgements of speech as formal or informal.*

This 3rd hypothesis is based largely on the assumption that hypothesis (2) is correct, but also on some previous work done on the relationship between prosody and the perception of politeness in both Japanese (Ito, 2001) and other languages (Laan, 1997). These studies (discussed further

in Chapter 3) both showed that there is a significant relationship between a number of prosodic variables and how listeners judge the formality of sentences.

(4) *Using the information from the experiments investigating hypotheses (1) – (3), a predictive Bayesian probabilistic statistical model will be able to accurately model the relationship between prosody and formality in spoken Japanese.*

The specifics of the statistical model will be discussed in greater detail in Chapter 5, but this hypothesis is largely based on the success of previous *ideal observer models* (Geisler, 2003) in predicting category membership under uncertainty based on knowledge of the distributions of phonetic parameters related to the category (Clayards et al., 2008; Kleinschmidt & Jaeger, 2015). Thereby, the objective will be to create a probabilistic model which is capable of classifying a recording of Japanese as either formal or informal based on prosodic information alone.

1.4 Summary of thesis content

1.4.1 Statistical analyses

Before describing the experiments that make up this thesis, a brief description of the statistical methods used in analyses will be useful. For the most part this thesis eschews the use of standard statistical techniques such as ANOVA, t-test, and linear regressions in favor of linear mixed effects regression models (Bates et al., 2015) for continuous data, and cumulative link mixed

models (Christensen, 2015) for ordinal data in R (R Core Team, 2016). The reasons behind using this approach are based largely on the design of the experiments in this thesis – one of the primary and most important assumptions of linear models (such as ANOVA) is that all observations are *independent* (Winter, 2013). In short, the independence assumption means that a linear model will assume that each data point has come from a different subject. For example, an experiment where you had 6 subjects, 3 male and 3 female, and measured the mean f_0 of each speaker in order to test the relationship between gender and mean f_0 would have independent observations (just one data point of the variable of f_0 per subject), and in such a case a linear model would be an appropriate statistical tool. However, none of the experiments that make up this thesis have independent observations of variables – there are a number of data points from each subject, and in a linear model this can results in within-group variance that will not be accounted for by the model. The addition of a *random effects structure* in a mixed effects model allows the model to calculate random intercepts and slopes for each subject (and for any other relevant factors, such as stimulus object, or time), which can explain more variance than a typical linear model, and therefore arrive at more accurate results (see e.g. Freeberg & Lucas, 2009; Lazic, 2010; Winter, 2011 for further discussion of this issue).

Although the models used throughout the thesis vary and will be described in their respective chapters, in general they follow the format shown in (5).

(5) *[variable] ~ [category] + [random effects]*

In (5), [variable] refers to the prosodic variable of interest, [category] refers to some binary category distinction, in this case often the formal/informal split, and the [random effects] consist of a random effects structure as justified by the experimental design, generally including the factor of subject. In plain language, the model can be read as [variable] as a function of [category] including unknown random variance. Additionally, the random slopes of the fixed factors of all linear mixed effects models presented are given in their entirety in Appendix I.

These models do not return P-values in the same way as linear models, but rather give likelihoods which express how well the fitted model accounts for the data. P-values are arrived at by comparing a full model to a 'null' model, which does not include the fixed effect of interest, in a likelihood ratio test (Casella & Berger, 2001). This test essentially compares how likely the data is under both models, and returns a p-value which expresses the probability of observing the data under the 'null' model.

Cumulative link mixed models vary somewhat from linear mixed effects models in the way they are designed in order to better deal with ordinal data. Rather than being coefficients of a linear regression, slope estimates for fixed factors in a *clmm* are instead coefficients on a normalized response scale indicating the most likely response for a given category. This allows for the models to be more easily interpreted in terms of the original ordinal scale.

1.4.2 Pilot study of the relationship between prosody and formality in Japanese

The second chapter of this thesis describes a pilot study which aims to conduct a preliminary investigation of research questions **(1)** and **(2)**. The experiment will attempt to elicit informal speech from subjects in the lab, by having them produce sentences containing geminate contractions (as described in Section 1.2.2.3). The relationship between prosody and formality will then be tested by having subjects produce a sentence that contains the singleton counterpart of the geminate contraction, but is otherwise identical, as in (6).

(6) a) *kono eiga wa* ***tsumaranai*** *desu*

This movie-TOPIC boring is "this movie is boring" (formal)

b) *kono eiga wa* ***tsumannai*** *desu*

This movie-TOPIC boring is "this movie is boring" (informal)

Based on previous work (Sagisaka & Miyatake, 1988; Ofuka et al, 2000), this experiment tests for significant co-variance of mean f_0, duration, and amplitude with the formal/informal contrast. Results show a significant relationship between both f_0 and duration with formality, with informal recordings exhibiting both a higher mean f_0, and shorter overall durations than their formal counterparts. These results provide initial support for hypotheses **(1)** and **(2)**, and better inform the corpus-based study that is the subject of Chapter 3.

1.4.3 Corpus based study of the prosody of informal conversational Japanese

The third chapter of this thesis describes a corpus-based study of spoken Japanese which addresses research questions (1) and (2) (described in Section 1.3.1), by investigating the prosodic properties of the informal register of speech via examination of recorded, conversational speech. Specifically, this study will examine the prosodic variables of mean f_0, articulation rate, and f_0 range (defined operationally as 4 SDs of f_0 in an utterance).

For this experiment, the goal is to make use of speech data that is both as natural as possible, and also conversational. The reason behind wanting to investigate conversational speech in particular is that the informal register of speech in Japanese (discussed in Section 1.2.2) that this thesis is concerned with occurs primarily in conversational situations, and therefore even corpora that contain spontaneous – but non-conversational – speech might not contain sufficient examples of informal speech to provide data for this study. Examination of available corpora of spontaneous Japanese shows that this is in fact the case. Because of the lack of appropriate data for this study, it will make use of a corpus of conversational Japanese created specifically for use in this project.

The corpus data will be annotated and segmented in order to fulfil the purposes of the study. Specifically, speech intervals from the subjects will be labelled, and the formality of each interval judged based on the observations regarding the properties of the formal and informal

registers of speech in Japanese that were discussed in Sections 1.2.2.1 and 1.2.2.2.

With the data labelled, mean f_0, articulation rate, and f_0 range will be analyzed using a combinations of mixed effects models (Described in Section 1.4.1) and a functional data analysis (Ramsay, 2006). These analyses show significant relationships between each prosodic variable and level of formality, where each is higher in informal than formal speech. The results show strong support for both hypotheses **(1)** and **(2);** it does appear that the formality of an utterance is related to its prosody on a number of levels, and that these changes in prosody are used consistently by speakers in the expression of formality in addition to any relevant lexical cues.

1.4.4 The effects of prosody on the perception of formality in delexicalized speech

The fourth chapter of this thesis is a perceptual study which investigates the salience of the statistical relationships observed in Chapter 3 to speech perception. The experiment addresses research question **(3)**, whether or not listeners make use of prosody when making judgements regarding the intended level of formality of speech. The hypothesis that listeners can in fact do so will be tested in an experiment via the manipulation of the prosody of synthetic de-lexicalized speech recordings.

In order to isolate the effects of prosody from the possible influence of any lexical cues, the experiment will use de-lexicalized speech as the stimuli. De-lexicalized speech refers to recordings which have had all of the lexical, syntactic, morphological, and phonological

information obscured, leaving only prosodic cues (Pagel et al., 1996). Such recordings are ideal for the purposes of this experiment, in that the only information they will provide to the listener is the prosody, allowing for a fair degree of confidence that there are no other variables in the recordings that are unduly influencing the results.

The delexicalized recordings to be used in this experiment will be created using a version of the Klatt synthesizer (Klatt, 1980; Klatt & Klatt, 1990; Iles & Ing-Simmons, 1994) parameterized by f_0 and amplitude data that will be measured from recordings of natural Japanese speech. In order to test the effect of differences in prosody on listeners' judgments of formality, the prosodic variables of mean f_0, articulation rate, and f_0 range will also be manipulated, and the differences in listener responses to manipulated and un-manipulated stimuli calculated and analyzed using mixed effects models.

Analysis of the data shows that there is a significant relationship between formality and changes in the prosody in speech perception, as predicted by hypothesis **(3).** It appears that listeners do use prosodic information when making judgements regarding formality, but the effect is only significant when a number of prosodic variables are manipulated together.

1.4.5 Modeling formality in Japanese using Bayesian inference

The final content chapter of this thesis describes a model under the Bayesian framework which makes use of information gathered in the experiments described in Chapters 3 and 4 to predict

whether a given recording is formal or informal based on the prosody alone. The model will make use of Bayes' rules (Bayes & Price, 1763) and Bayesian inference (Box & Tiao, 2011) to calculate the probability of membership in either the formal or informal category, given X, while taking a degree of uncertainty into account.

The model will make use of Bayesian logistic mixed effects regressions to calculate weights for each predictive variable (mean f_0, articulation rate, and f_0 range), and then use those weights to calculate both prior probabilities and likelihoods for any given recording of spoken Japanese, which can then be used to calculate the posterior probability of formality for that recording.

In total, three different versions of the predictive model will be tested – a version based solely on the prior probabilities, a version that includes the likelihoods, and a version that uses a different prior specification to counteract problems with model overfitting. The first two models are fairly accurate, able to correctly predict formal speech ~60% of the time and informal speech ~67% of the time in the case of the priors-only model, and formal speech ~63% of the time and formal speech ~74% of the time in the case of the priors and likelihood model. The model using the different prior specifications does not suffer from an imbalance in accuracy when predicting the two categories, and is able to predict formal speech ~71% of the time, and informal speech ~68% of the time.

The model is not perfect, but given the challenge that listeners themselves experienced in the study in Chapter 4 in categorizing recordings as formal or informal based on delexicalized prosody, it is a very positive sign that it is possible to create a model that predicts at a rate ~20% better than chance. The model performs better than human listeners in the experiment described in Section 1.4.4 at predicting formality given only prosodic information, and appears to be an accurate and appropriate means of modeling a categorization task in speech perception.

Chapter 2

Pilot Study of the Prosodic Properties of Formality in Japanese

2.1 Introduction

2.1.1 Chapter overview

Before embarking on a full-scale investigation of the role of prosody in the expression of formality in spoken Japanese, it is helpful to first develop a basic picture of this relationship to build from. While previous studies of the relationship between prosody and formality in Japanese have shown some significant results (see Ofuka et al., 2000; Ito, 2002), these results are not consistent (i.e. Ofuka et al., 2000 shows a relationship where informal speech has a significantly higher mean f_0, while Ito, 2002 shows the opposite results). In order to decide what variables will be worth investigating in larger scale studies, a smaller scale pilot study focused on a single condition was conducted. This condition was **gemination**, and more specifically the phonological phenomenon where geminates are used as contractions in combination with elision in spoken Japanese, as in /dokoka/ ("somewhere") → [dok:a] (see Arai, 1999; Kawahara, 2015 for brief discussions). These geminate contractions appear almost exclusively in an informal, conversational register of spoken Japanese (Arai, 1999), and were therefore judged to be a good method of eliciting informal speech from subjects in a lab. Geminates are also useful for testing the binary formal/informal categories as their singleton counterparts provide an analogous category split which makes the construction of experimental stimuli fairly straightforward (see Section 2.2 for further details).

This chapter will first give a brief description of geminate contractions in Japanese from both a phonological and pragmatic perspective, and will also review previous work on the acoustic correlates of gemination in Japanese more generally to help to better inform the experimental design of the pilot. Further sections will describe the experimental design and the recordings produced, and how these were analyzed in order to provide a basis for the study in Chapter 3.

2.1.2 Geminate contractions in Japanese

The primary treatment condition of the experiment described in this chapter involves eliciting informal speech via the production of geminate contractions. The term geminate contractions here refers to instances in spoken Japanese where words are shortened via elision (generally of a vowel but also sometimes of VV sequences spanning two moras) (Kawahara 2015), and the following phoneme is then lengthened into a geminate segment. Examples of this phenomenon can be seen in (1).

(1) a) /wakaɻanai/ ("don't understand") → [wakan:ai]

b) /dokoka/ ("somewhere") → [dok:a]

c) /atatakai/ ("warm") → [at:akai]

d) /to jɯ ka/ ("by the way...") → [tsɯ:ka]

e) /oɻe no ɯt͡ʃi/ ("my house") → [orentʃ:i]

f) /ɯɻɯse:/ ("shut up") → [ɯs:e:] → [s:e:]

(1) shows examples of this phenomenon occurring in a number of phonological conditions, including instances of elision word-internally, word initially (as in 1f), across multiple words, and even in cases where the resulting long nasal consonant (in example 1a) would violate the standard rules of Japanese phonology prohibiting voiced geminates (Mester & Ito, 1995). It is debatable whether these geminate contractions should be treated as phonetically and phonologically identical to other (lexical) geminates in Japanese, or if it simply a matter of the elision of a single vowel combined with assimilation of the consonant cluster into one longer sound (i.e. in such a case 1a would actually be more accurately represented as /wakarnai/ → [wakan:ai]). The acoustic or articulatory analysis required to decide one way or another is beyond the scope of this thesis, but would be an interesting topic for future research.

The conditions which allow these geminate contractions to occur are somewhat opaque, and are not widely studied in the phonological or sociolinguistic literature of Japanese. Although there do appear to be some commonalities among the phonological conditions where elision **can** occur in Japanese – for example, in sequences where consecutive moras contain identical consonants or vowels, or in cases where multiple vowels occur in succession across words – there does not appear to be any particular condition where such elision **must** occur. This leads to the conclusion that geminate contractions in Japanese are a gradient phenomenon, the occurrence of which is likely affected by pragmatics and the social context of a given utterance.

One observation that has been made is that geminate contractions mainly occur in particular varieties of spontaneous speech – Arai (1999) describes how in a corpus of spontaneous Japanese, gemination applies "in more words and more environments as a type of fast speech variation than it does as a phonological process" (Arai, 1999: 616). Although what exactly is meant by 'fast speech' is left somewhat ambiguous, it is contrasted with a more 'careful' style of speech; these two speech styles could potentially correlate with informal and formal speech styles respectively. This possibility is supported by observational evidence of spoken Japanese – where geminate contractions have been observed as mainly occurring in informal, conversational speech – and led to the hypothesis that eliciting these geminate contractions from subjects in an experiment would cause them to produce utterances with the prosodic properties of informal speech.

Due to the fact that the geminate/singleton contrast will be used in this study as an index of the formality of an utterance, it is important to determine whether or not there are any acoustic correlates of gemination itself which could cause confounds, wherein apparent relationships between acoustic variables and formality are actually caused by the relationship between the variables of interest and the presence/absence of the geminate segments themselves. To this end, in order to determine how best to treat the geminate segments in the experimental design, the following section will present a more general review of previous work on the acoustic properties

of gemination in Japanese.

2.1.3 Literature review

One of the first studies of the acoustics of gemination in Japanese was Smith (1993). In that study, the articulatory and acoustic properties of gemination in Japanese and Italian were compared. The experimental design for the sections regarding Japanese was somewhat questionable, as it used target utterances that are phonologically prohibited in Japanese (/-ti/ and /-t:i/ and /-m:i/), and which also violate the general rule in Japanese words against having multiple [+voice] stops in a single word. Nevertheless, it did reveal a number of the articulatory properties of gemination in Japanese, which became clearer when compared with another language.

Firstly, stop closure for geminates in Japanese was different than in Italian from a timing perspective. While in both languages the closure duration of the geminate segment was clearly longer than a corresponding singleton, in Japanese the closure of the singleton stop occurred earlier than for the geminate, while in Italian the closure of the singleton fell right in the middle of where the closure occurred for the geminate (Smith, 1993: 49-50). This was caused by durational differences in the vowels preceding the geminate segment; in Japanese, the vowel preceding the geminate segment was longer, while in Italian the vowel preceding the singleton was longer. Differences in timing are not in and of themselves tremendously surprising, as Japanese is mora-timed while Italian is (aside from certain dialects) syllable-timed (Payne, 2005).

However, it is somewhat unexpected, due to previous studies showing a consistently shorter duration of the vowels preceding geminates in languages such as Malayalam, Italian, and Tashlhiyt (Local & Simpson, 1999; Payne, 2005; Ridouane, 2007), and hence an earlier closure. This finding indicates that gemination in Japanese might in some ways behave counter to previously defined expectations.

A few more properties of gemination in Japanese were found in Smith (1993). Notably, both the position of the tongue body and the lip aperture was found to vary systemically based on the geminate-singleton contrast (Smith, 1993: 49), indicating that there are likely further acoustic correlates of gemination in Japanese caused by these articulatory differences. These further acoustic correlates are of interest to the current study, as they could potentially confound our target variables. Although at the time Smith (1993) made the claim that duration was the only parameter that varied with the geminate-singleton contrast, further research revealed that not to be the case. The first such study was Kawahara (2006), which examined the specific phenomenon of geminate devoicing in borrowed words in Japanese from the perspective of Optimality Theory (Prince & Smolensky, 2002), a system of phonological computation where a language's surface forms are determined by testing possible outputs against a series of ranked phonological constraints. Although the motivation for the study was phonological, it was conducted via an experimental phonetic study of the acoustic properties of gemination in Japanese.

In addition to the devoicing of geminates – an unsurprising phenomenon due to the general rule against voiced geminates in native Japanese – a number of other non-durational acoustic correlates of gemination in Japanese were observed. The first such finding was the fact that there were changes in f_0 in the sounds surrounding geminate consonants. Although this was somewhat glossed over in Kawahara's study due to it being unrelated to the voicing contrast the study was focused on, it was observed that f_0 was significantly higher in the vowel before a geminate, and fell more rapidly than expected (when it fell due to pitch accent) after the geminate (Kawahara, 2006: 558). There are a few possible acoustic factors that could lead to this change in f_0, but a logical explanation, posited by Kawahara, is that the change is greater simply because the speaker has more time to make the change due to the increased closure duration of geminates. This explanation appears reasonable, although it does not entirely account for the increased f_0 before geminates.

Another finding of note, echoing the observations regarding the timing of closure found in Smith (1993) was the fact that vowels preceding geminates were significantly shorter than those preceding singletons. Vowels were on average ~20ms shorter before geminates than singletons ($p < .001$). This is a rather striking difference between Japanese and other languages, although it was not unexpected due to previous findings. This significant difference in the durations of nearby segments based on the geminate-singleton contrast is important to the current

study, as any durational measurements relative to an utterance's formality could be unduly influenced by the presence of a geminate segment. This potential confound and methods of addressing it will be discussed further in Section 2.2. Although Kawahara (2006) was not specifically focused on discovering acoustic correlates of gemination, further studies were based on its findings.

The study which has done the most in-depth examination of the acoustic correlates of gemination in Japanese was Guion & Idemaru (2008). Based upon previous studies of gemination, they investigated the potential covariance of the geminate-singleton contrast with acoustic factors such as intensity, voice quality, and fundamental frequency, in addition to any durational correlates. The idea that intensity is correlated with gemination was based on a study on Malay (Abramson, 1997) where consonant amplitude covaried with the presence or absence of gemination. A production study of Japanese was conducted, using a set of 36 minimal pairs which varied only in the geminate/singleton contrast. These words were pronounced in carrier sentences, and were all constructed using the same phonological pattern (/seCa/ where C is some consonant). The majority of these words were nonce-words, but none were ill-formed based on other phonological rules. This mixing of real and nonce forms was not considered ideal by Guion & Idemaru, but it is virtually impossible to avoid when attempting to create novel forms in Japanese while following a pattern, largely due to the restrictions of the possible moraic

structures. Due to this, the mixing was accepted as a possible confound.

In terms of durational findings, in addition to replicating previous findings of a vowel preceding a geminate being significantly longer, it was also found that the vowel following a geminate segment would exhibit the opposite effect, and would be significantly shorter than after a singleton. The length was also affected by the voicing of the geminate, but the results were consistent regardless of voicing. This finding indicated that possible co-articulatory effects could occur post- as well as pre-geminate, which will have some bearing on the experimental design of this chapter.

The main findings of significance in Guion & Idemaru (2010), however, related to non-durational correlates of gemination. All of the effects tested for – intensity, f_0, and amplitude – showed a significant difference between singletons and geminates. Each was measured from the start of the vowel preceding the geminate/singleton to the end of the vowel following. Intensity was uniformly greater in the vowel preceding a geminate than in the vowel following, while the opposite was true for singletons. This difference was statistically significant ($p < .0001$). f_0, similarly, showed the expected effect (Kawahara, 2006) of dropping much more sharply after a geminate than after a singleton, but was also found to be higher overall in geminate segments than in singletons. The difference in the shift in f_0 was significant ($p < .0001$). Amplitude showed the least co-variation with gemination, though there was a greater fall in amplitude from the first

harmonic of the preceding vowel to the first formant peak before a geminate consonant. The fact that these acoustic variables have a significant relationship with the geminate-singleton contrast will be taken into account in the analysis of such variables in relation to formality in Japanese, particularly when analyzing f_0, which is a primary variable of interest in this study.

2.2 Production study

2.2.1 Research questions and hypotheses

The primary research question investigated by this study is *which, if any, acoustic variables have a significant relationship with the level of formality of an utterance in spoken Japanese*? Although previous work on the relationship between prosody and the expression of formality in Japanese (Ofuka et al., 2000; Ito, 2002) and in spoken language in general (Hidalgo Navarro & Cabedo Nebot, 2014) has shown that there are a number of variables – such as f_0, articulation rate, amplitude, and vowel quality – which have been observed as having a significant relationship with formality, these results are not necessarily consistent among different languages, or even within a single language (as in the contrasting results of Ofuka et al., 2000 and Ito, 2002). For this reason, this pilot study seeks to investigate this question with an eye toward determining which variables should be of interest for the larger-scale corpus-based study to be described in Chapter 3.

The hypothesis that the experimental portion of this chapter is testing is that there will be a number of acoustic co-variates with the level of formality of an utterance in spoken Japanese. Three of potential co-variates will be tested for based on previous work: mean f_0, duration, and amplitude. Based on previous results, my initial hypothesis is that all of these variables will be significantly higher in informal than in formal utterances. These variables have all been found to co-vary significantly with the geminate/singleton contrast, and so care must be taken in the analysis to avoid any potential false positives.

2.2.2 Experimental design

The design of the experiment was based upon Guion & Idemaru (2008). Minimal pairs or nearly minimal pairs representing the singleton/geminate contrast were used as stimuli in a phonetic production experiment, which was conducted in the University of Oxford Phonetics Laboratory in November-December 2012. These minimal pairs were placed in carrier sentences and displayed to subjects to be read aloud. Although the design was similar to Guion & Idemaru (2008), most of the details of the experiment conducted for this study were markedly different.

The main difference was that the minimal pairs used were all real words, and each was placed in a natural, semantically sensible carrier sentence rather than a generic meaningless one. This was done in order to attempt to represent a stylistically natural condition where the target geminates of interest might occur in speech. Previous lab-based studies (Guion, 1995) have shown

results where significant effects were found when words were produced in natural sentences, but were absent when the same words were produced in generic carriers. Therefore, sentences were created, with the help of native speakers of Japanese, which could contain both words of the minimal pair in identical position within the sentence while still remaining semantically natural, thereby resulting in full minimal pair sentences where the only phonological difference was the geminate/singleton contrast. In total 14 pairs of words were used, meaning there were 28 target stimuli in total. A full list of all of the minimal pairs used (broken down into treatment and control groups), as well as a full list of the carrier sentences used, can be seen in Tables 2.1 and 2.2 respectively.

Table 2.1: *List of experimental stimuli minimal pairs, including glosses, broken into treatment and control conditions. * Regarding sentence 12, in Japanese phonology, an underlying /h/ will always be geminated into /p:/ post-lexically, as in /haʧi/ ("eight") + /hjakɯ/ → [hap:jakɯ] ("eight hundred").*

Treatment Condition				
Pair #	**Singleton**	**Gloss**	**Geminate**	**Gloss**
1	/wakaɻanai/	'don't understand'	[wakan:ai]	'don't understand'
2	/tsɯ̟maɻanai/	'boring'	[tsɯ̟man:ai]	'boring'
3	/so:ka/	'I see'	/so:k:a/	'I see'
4	/atatakakɯnai/	'not warm'	/at:akakɯnai/	'not warm'
5	/dokoka/	'somewhere'	/dok:a/	'somewhere'
Control Condition				
6	/omae/.../ʃi̟teɻɯ/	'to do' (forceful)	/omae/.../ʃi̟t:eɻɯ/	'to know' (forceful)
7	/kimi/.../ʃi̟teɻɯ/	'to do' (casual)	/kimi/.../ʃi̟t:eɻɯ/	'to know' (casual)
8	/mate/	'wait' (imperative)	/mat:e/	'waiting'
9	/masatsɯ̟ /	'to rub'	/mas:atsɯ̟ /	'to erase'
10	/saki/	'ahead'	/sak:i/	'before'
11	/kateɻɯ/	'can win'	/kat:eɻɯ/	'to be winning'
12	/jahari/	'after all'	[yap:ari]	'after all'
13	/to/	QUOT-particle	[t:e]	QUOT-particle
14	/ita/	'was at a place'	/it:a/	'went'

Table 2.2: *List of sentence pairs used as experimental stimuli, along with translations. The alternating singleton/geminate pairs are within {}.*

Carrier Sentence (Romanized/non-IPA)	Gloss
'wake ga {wakaranai/wakannai} yo.'	I {don't understand} the reason.
'kono eiga wa {tsumarainai/tsumannai} ne.'	This movie {is boring}, huh.
'aa, {souka/sokka}'	Ah, {I see}.
'kono gohan wa {atatakakunai/attakakunai}.'	This rice/food {isn't warm}.
'aitsu o {dokoka/dokka} de mita koto aru.'	I've seen him {somewhere} before.
'omae wa nani o {shiteru/shitteru}?'	What {are you doing/do you know}? (forceful)
'kimi wa nani o {shiteru/shitteru} no?'	What {are you doing/do you know}? (casual)
'soko de {mate/matte}.'	{Wait/Please wait} over there.
'ano mono wa {masatsu/massatsu}shita.'	(I/he/she/you) {rubbed/erased} that already.
'shokudou e {saki/sakki} ikeba yokatta.'	(I/he/she/you) should have gone to the diner {ahead of someone/earlier}.
'aitsu ni wa {kateru/katteru} yo.'	I {can beat him/am beating him}.
'sore wa {yahari/yappari} ichiban benri da.'	That's the most convenient {after all}.
'sou da {to/tte} itta yo ne.'	(I/he/she/you) {QUOT} said so, right?
'soko ni {ita/itta}.'	(I/he/she/you) {was/went} there.

The pairs were also broken down into treatment and control groups in order to be as certain as possible that there were no confounds being brought about by the geminates themselves. The treatment category consisted of the sentence pairs containin geminate contractions (for example /at:akakɯnai/ and /atatakakɯnai/). The control condition was made up of sentence pairs containing either post-lexical geminates where the pair was made up of two possible realizations of the same word, as in /jahaɾi/ and /jap:aɾi/ ("after all"), and lexical geminates where the pair

was made of two different words, as in /ita/ ("was at a place") and /it:a/ ("went").

There are some potential confounds related to the experimental stimuli. Foremost among these is the issue of the lemma frequency of the geminate/singleton pairs used. Lemma frequency refers to the semantic frequency of a word, or more simply how common a given word is in a language. Previous work (Gahl, 2008; Bybee, 2001, 2006) has shown that words with higher lemma frequency in English are produced with consistently shorter durations, indicating that lemma frequency is likely an important factor in language production. As this experiment was a production study with a focus on durational measurements, controlling the lemma frequencies of the stimuli was important.

In order to account for this confound, geminate/singleton pairs with the highest possible lemma frequencies were chosen. Frequency measurements were based on the Internet Corpus of Japanese, which is a compilation of writing from various sources taken from the internet, containing roughly 253 million lemma tokens (http://corpus.leeds.ac.uk/frqc/internet-jp.num). Frequency measurements for the corpus were given in parts per million, and all of the words chosen had frequencies of at least 20 parts per million. The majority of words had a frequency of over 100 parts per million, which in terms of the corpus examined placed them among the 1,000 most common words in the language. This is not a perfect way to measure word frequency as the measurements are based on a corpus of written data rather than spoken, but the extremely large

sample size should at the very least provide a reasonable estimate of a given word's frequency.

Five subjects were recruited to participate in the experiment (three females, two males). All were native speakers of Japanese aged from 23 to 33. There are some possible confounds related to the subjects. As measurement of f_0 was a major part of this experiment, it was important that all subjects spoke a single variety of Japanese with a consistent pitch-accent system. There is a large amount of variability in the pitch accent systems of Japanese (which is the realization of f_0 in speech) in different dialects (Kubozono, 2012) which could result in f_0 measurements that are drastically different from those in standard Japanese. In addition to the issue of dialect, it has been shown in previous studies (Munson et al, 2011) that subjects with previous experience in phonetics tend to have skewed experimental results, increasing the possibility of a Type I or Type II statistical error. A final potential confound relating to the subjects was age. Studies of the vocal apparatus and f_0 production over many years (Harrington et al, 2007; Mwangi et al, 2009) have shown that mean f_0 declines with age, which could present a problem for the accuracy of the statistical measures in this experiment if a subjects' f_0 is affected by their age.

In order to, as much as possible, eliminate the confounds and biases relating to the subjects, each potential volunteer was asked to answer a brief questionnaire in order to determine their suitability for the experiment. In order to eliminate any issues related to dialects, only subjects who grew up speaking the Tokyo dialect of Japanese were accepted. All subjects were also

queried regarding any previous training in linguistics, specifically phonetics and phonology. No subjects reported any previous training. Finally, all subjects were aged from 23 to 33; by minimizing the age range of subjects, the aim was to also minimize the potential variability of their fundamental frequency production due to age, as well as to minimize any stylistic or sociolinguistic differences due to age.

In the experiment, three different pseudo-random orderings of stimuli were used, where the order of the sentences to be presented was initially randomized, and then manually changed to avoid any sentences containing the same geminate/singleton target pair occurring successively. In addition to the target sentences, there were 28 separate distractor sentences mixed randomly among the targets. This random ordering of stimuli appeared to be effective, as in spite of the presence of minimal pair sentences, none of the subjects suspected the formal/informal contrast as the target of the experiment when queried after the experiment. Subjects were shown the stimuli on a computer screen, and were asked to read the entire sentence before pronouncing it as naturally as possible. Each sentence (including the distractors) was presented a total of six times during the experiment. Subjects were allowed to read the sentences and pronounce them at their own pace, and each session took from 25 to 30 minutes. Across all the subjects, this resulted in 840 total recordings of the target sentences.

All stimuli were presented entirely in hiragana orthography, in order to avoid any potential

conflicts with the speed of processing different scripts. Previous research using fMRI has shown that the different Japanese orthographic scripts activate different areas of the brain when they are read (Nakamura et al, 2005), and so the decision was made to use only one script in order to avoid any potential confounds related to visual language processing.

Analysis of the recorded data was done in Praat (Boersma & Weenink, 2017). Each recorded sentence was segmented in a Praat Text Grid into intervals made up of the portions of the utterance surrounding but not including the geminate/singleton word pair, and the word pair itself. An example of this segmentation of a recording in shown in Figure 2.1. Durational measurements of each interval were taken, and means of f_0 and amplitude (dB) were taken from the portions of the utterance not including the geminate/singleton word pair (i.e. from the circled "pre" and "post" labels seen in Figure 2.1, but not from the "gem" label).

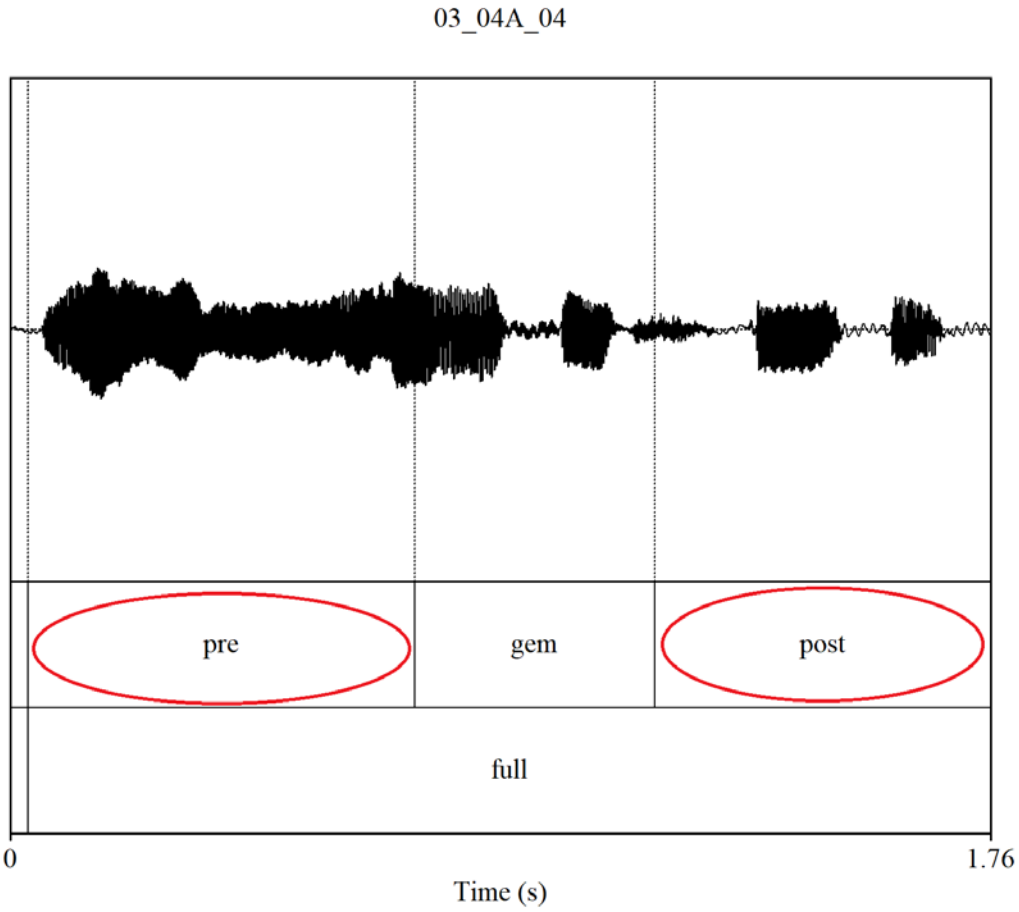


Figure 2.1: *Example of a waveform and labelled text grid for one of the recordings made in this study. The label "gem" indicates the word containing the geminate segment. Measurements were taken only for the circled portions of the text grid.*

The reasoning behind separating the geminate/singleton pair from the rest of the utterance in the analysis was that previous studies (Kawahara, 2006; Guion & Idemaru, 2008) have found that the presence of geminate segments has an influence on the prosody of the surrounding sounds, and of the geminate segments themselves. Since the supra-segmental features of the utterance (f_0 and duration in this case) are some of the main variables being analyzed in this study, leaving out the words containing the geminate segments was judged to be the best way to limit any potential

confounds. Combined with the control groups containing lexical and post-lexical geminates for comparison, this should fulfill the goal of minimizing the effects gemination may have on the prosody of the recordings made during the experiment, and allow us to investigate the relationship between prosody and formality.

2.3 Data and analysis

2.3.1 Data overview

All recordings were examined manually for interference, or production errors on the part of the subjects. All recordings appeared free of interference and presented clean waveforms and smooth pitch tracking contours. There were two instances of obvious production errors by the subjects where the wrong word was pronounced. These instances were excluded from the analysis. Mean f_0 and amplitude were calculated using Praat scripts for the portion of each utterance not including the geminate/singleton pair. Table 2.3 shows overall descriptive statistics for the variables of interest.

Table 2.3: *Descriptive statistics for the three tested variables.*

Variable	Mean	Standard Deviation
Mean f_0	226.2 Hz	60.9 Hz (27% of mean)
Duration	540 ms	145 ms (27% of mean)
Amplitude	72.4 dB	4.9 dB (7% of mean)

Although these descriptives do not tell us a huge amount about the details of the data, we can see

that the standard deviations are somewhat high for mean f_0 and duration, while intensity appears fairly consistent. Examining the means and SDs in greater detail provides some insight about the patterns of variation occurring within the data. Table 2.4 shows a breakdown of the means and SDs of the variables for each of the 5 subjects in the experiment, while Table 2.5 shows the variables broken down by geminate/singleton sentence pair.

Table 2.4: *Summary of means and SDs of each variable by subject.*

	f_0		Duration		Intensity	
Subject	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
1	177.1 Hz	12.2 Hz (7%)	482 ms	117 ms (24%)	79.6 dB	2.6 dB (3%)
2	289.8 Hz	29.1 Hz (10%)	617 ms	160 ms (26%)	70.1 dB	3.2 dB (5%)
3	263.1 Hz	17.7 Hz (7%)	530 ms	129 ms (24%)	69.1 dB	3.1 dB (4%)
4	139.6 Hz	19.8 Hz (14%)	487 ms	119 ms (24%)	73.8 dB	3.1 dB (4%)
5	261.7 Hz	17.1 Hz (7%)	586 ms	145 ms (24%)	69.7 dB	2.7 dB (4%)

Table 2.5: *Summary of means and SDs of each variable by sentence pair.*

	f_0		Duration		Intensity	
Sentence	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
1	230.1 Hz	64.0 Hz (28%)	430 ms	65 ms (15%)	71.7 dB	4.8 dB (7%)
2	223.0 Hz	57.1 Hz (26%)	627 ms	76 ms (12%)	74.4 dB	3.9 dB (5%)
3	217.2 Hz	62.3 Hz (29%)	450 ms	67 ms (15%)	70.6 dB	4.1 dB (6%)
4	224.3 Hz	52.5 Hz (23%)	725 ms	82 ms (11%)	74.5 dB	4.2 dB (6%)
5	230.2 Hz	54.6 Hz (24%)	450 ms	64 ms (14%)	70.6 dB	4.7 dB (7%)
6	228.5 Hz	63.2 Hz (28%)	744 ms	100 ms (13%)	76.0 dB	5.2 dB (7%)
7	242.6 Hz	60.0 Hz (25%)	690 ms	95 ms (14%)	74.3 dB	4.4 dB (6%)
8	217.1 Hz	58.1 Hz (27%)	480 ms	75ms (16%)	71.1 dB	4.6 dB (6%)
9	223.7 Hz	55.8 Hz (25%)	625 ms	134 ms (21%)	75.9 dB	4.9 dB (6%)
10	230.7 Hz	57.8 Hz (25%)	595 ms	80 ms (13%)	71.1 dB	4.4 dB (6%)
11	227.8 Hz	63.4 Hz (28%)	541 ms	66 ms (12%)	72.1 dB	5.4 dB (7%)
12	221.5 Hz	60.9 Hz (27%)	406 ms	56 ms (14%)	71.2 dB	4.1 dB (6%)
13	242.7 Hz	73.8 Hz (30%)	369 ms	49 ms (13%)	70.4 dB	4.4 dB (6%)
14	208.1 Hz	63.5 Hz (31%)	446 ms	61 ms (14%)	70.5 dB	4.0 dB (6%)

These tables show that there is a noticeable difference in both the amount of variation (SD) in some of the variables of interest (mean f_0 and duration) and in the means of the variables, based both potentially on the which subject is speaking (Table 2.4), and which sentence is being spoken (Table 2.5). For example, speaker 4 shows twice as much variation in their f_0 as compared to three other speakers, and some of the speakers (2 and 5) seem to articulate much more slowly than the others. This, along with the fact that the design of the experiment means that the observations are not independent, indicates that a standard ANOVA is likely not an appropriate

tool for analyzing the data in this experiment. Computationally, an ANOVA computes and compares the variance in a variable between categories (in the case of this experiment, the geminate and singleton categories). However, Tables 2.4 and 2.5 indicate that there is additional unpredictable variation both between the different speakers, and the different sentence pairs. In order to provide a sufficiently robust random effects structure for the analysis, mixed effects regression models will be used as a tool for statistical analysis in the following sections. These models will test each prosodic variable as a dependent variable, with a fixed factor of *formality* and subject *gender* for the treatment condition. The random effects structure will test random intercepts and slopes (of *formality*) for each subject and sentence pair. A separate model will also be run on the full data set (both treatment and control conditions) which will additionally test for a possible interaction with *condition* (i.e. testing if any significant differences in the treatment condition appear to be driven by differences between the treatment and control conditions, which is what is hypothesized).

A more in-depth examination of the variables based on both the geminate/singleton contrast and on the experiment's control and treatment groups shows evidence of a difference between both the types of geminate/singleton sentence pairs analyzed, and between sentences containing either singletons or geminates in the treatment condition. To review (Section 2.1), based on the hypothesis that sentences containing geminate contractions would be more informal

utterances, the treatment condition of the experiment was singleton/geminate sentence pairs containing geminate contractions, while the control condition was sentence pairs containing only lexical geminates. Further, the hypothesis being tested was that the prosodic variables mean f_0, duration, and intensity would co-vary significantly with the formal/informal contrast. Table 2.6 shows a summary of the variables of interest broken down both by control/treatment condition, and by the geminate/singleton contrast.

Table 2.6: *Mean values of the target variables broken down by control/treatment condition and geminate/singleton contrast.*

	Control		Treatment	
Variable	**Singleton**	**Geminate**	**Singleton**	**Geminate**
Mean f_0	227.2 Hz	226.8 Hz	217.2 Hz	232.8 Hz
Duration	551.26 ms	536.85 ms	549.09 ms	519.88 ms
Mean Intensity	72.5 dB	72.5 dB	72.4 dB	72.3 dB

2.3.2 Intensity

One point that is immediately apparent from Table 6 is that there is very little variance in mean intensity between the categories. In fact, it appears that speakers produced sentences with very nearly identical intensities across all categories, and mixed effects models (shown in 2a and 2b) indicate that there is no relationship between intensity and either category or the geminate/singleton contrast. Intercept and slope coefficients for the random effects in these models can be found in Appendix 1.1.

(2) a). *Full Data Interaction Model*

Full Model: y = Mean Intensity, β = Formality : Condition + gender,

u = Intercept + Slope of Formality | Speaker and Sentence Pair

Null Model: y = Mean Intensity, β = Formality + Condition + gender,

u = Intercept + Slope of Formality | Speaker and Sentence Pair

Coefficients: FormalityF : ConditionT = 0.05 ± 0.32, t-value = 0.17

FormalityF = -0.04 ± 0.20, t-value = -0.22

ConditionT = -0.18 ± 1.21, t-value = -0.15

GenderM = 6.51 ± 1.66, t-value = 3.92

Random Effects:

Groups	Name	Variance	Std.Dev.	Corr
pair	(Intercept)	4.55172	2.1335	
	FormalityF	0.02752	0.1659	-1.00
speaker	(Intercept)	3.42188	1.8498	
	FormalityF	0.01132	0.1064	1.00
Residual		4.78192	2.1868	

Model Comparison Results: $DF(X) = 1$, $X^2 = 0.02$, $Pr(>X^2) = .86$

b). *Treatment Model*

Full Model: y = Mean Intensity, β = Formality : gender,

u = Intercept + Slope of Formality | Speaker and Sentence Pair

Null Model: y = Mean Intensity, β = gender,

u = Intercept + Slope of Formality | Speaker and Sentence Pair

Coefficients: FormalityF : GenderM = 0.24 ± 0.50, t-value = 0.48

FormalityF = -0.08 ± 0.31, t-value = -0.27

GenderM = 6.62 ± 1.75, t-value = 3.77

Random Effects:

Groups	Name	Variance	Std.Dev.	Corr
speaker	(Intercept)	3.5631527	1.88763	

	FormalityF	0.0249854	0.15807	1.00
pair	(Intercept)	3.3762036	1.83744	
	FormalityF	0.0008304	0.02882	-1.00
Residual		4.1587455	2.03930	

Model Comparison Results: $DF(X) = 2$, $X^2 = 0.23$, $Pr(>X^2) = .89$

2.3.3 f_0

The first variable of interest is mean f_0. The descriptive statistics discussed in the previous section indicated that there is a possible relationship between f_0 and the geminate/singleton contrast in the treatment condition. Figure 2.2 shows a bar graph visualizing the relationship between formality and f_0.

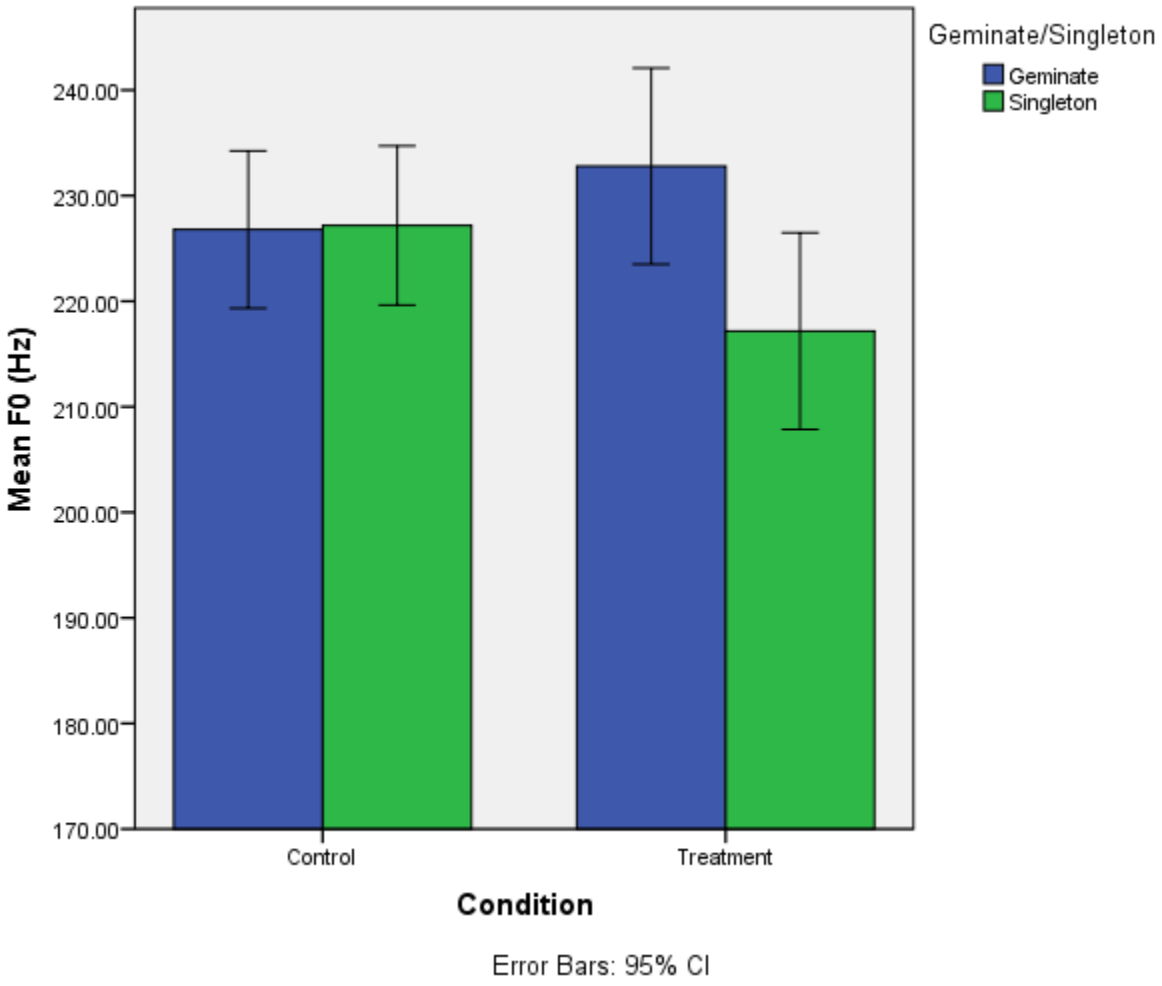


Figure 2.2: *Bar graph showing the relationship between mean f_0 and the geminate/singleton contrast in the portion of the utterance excluding the singleton/geminate segment in both experimental conditions. Error bars indicate the 95% confidence interval.*

It is fairly visually apparent that there is a difference in mean f_0 based on the geminate/singleton contrast only in the treatment condition, indicating that the difference could be related to the level of formality of the sentence rather than to the geminate/singleton contrast itself. The lowered mean f_0 in the sentences containing singletons in the treatment condition as compared to the control indicates an interaction, and the significance of this interaction is confirmed by the mixed

model in (3a).

Analysis of the relationship between mean f_0 and the geminate/singleton contrast in the treatment condition using model comparison of mixed effects regression models (Bates et al., 2015) (shown in (3b)) in R (R Core Team, 2017) shows that the relationship is significant. A further summary of all modelling results for this chapter can be seen in Table 2.7. The models in (3) include gender as a fixed factor as it has a predictable relationship with f_0, and also include an interaction between formality and gender to test if any relationship between f_0 and formality are driven by gender-based differences. Random slopes are included for the random effects of speaker, and sentence pair. The coefficients of these random slopes show that although there is some variation among speakers and sentence pairs the overall main effect of the geminate/singleton contrast remains consistent. The interaction between formality and condition (treatment or control) was tested in (3a) to determine if there are significant differences between the conditions, as hypothesized.

(3) a). ***Full Data Interaction Model***

Full Model: y = Mean f_0, β = Formality : Condition + gender,

u = Intercept + Slope of Formality | Speaker and Sentence Pair

Null Model: y = Mean f_0, β = Formality + Condition + gender,

u = Intercept + Slope of Formality | Speaker and Sentence Pair

Coefficients: FormalityF : ConditionT = -15.55 ± 2.95, t-value = -5.26

FormalityF = 0.72 ± 1.77, t-value = 0.41

ConditionT = 6.33 ± 5.63, t-value = 1.12

GenderM = -111.98 ± 14.25, t-value = -7.85

Random Effects: Groups	Name	Variance	Std.Dev.	Corr
pair	(Intercept)	92.05009	9.5949	
	FormalityF	8.10655	2.8472	-0.44
speaker	(Intercept)	240.45359	15.5066	
	FormalityF	0.08794	0.2965	1.00
Residual		298.75387	17.2845	

Model Comparison Results: $DF(X) = 1, X^2 = 15.325, Pr(>X^2) < .001$

b). *Treatment Model*

Full Model: y = Mean f_0, β = Formality : gender,

u = Intercept + Slope of Formality | Speaker and Sentence Pair

Null Model: y = Mean f_0, β = gender,

u = Intercept + Slope of Formality | Speaker and Sentence Pair

Coefficients: Formality : Gender = -0.85 ± 3.83, t-value = -0.22

Formality = -14.56 ± 2.78, t-value = -5.23

Gender = -106.66 ± 15.51, t-value = -6.87

Random Effects: Groups	Name	Variance	Std.Dev.	Corr
speaker	(Intercept)	280.0341	16.734	
	FormalityF	0.1998	0.447	-1.00
pair	(Intercept)	40.4815	6.363	
	FormalityF	8.9909	2.998	-0.92
Residual		263.0551	16.219	

Model Comparison Results: $DF(X) = 2, X^2 = 11.253, Pr(>X^2) = .003$

2.3.4 Duration

The final variable analyzed in this experiment was the overall durations of the sentence pairs not including the geminate/singleton segments themselves. As the carrier sentences should be

otherwise identical with these segments removed, duration can also serve as a basis for measuring differences in speech rate between the conditions. Figure 2.3 shows a bar graph visualizing the relationship between duration and the geminate/singleton contrast across conditions.

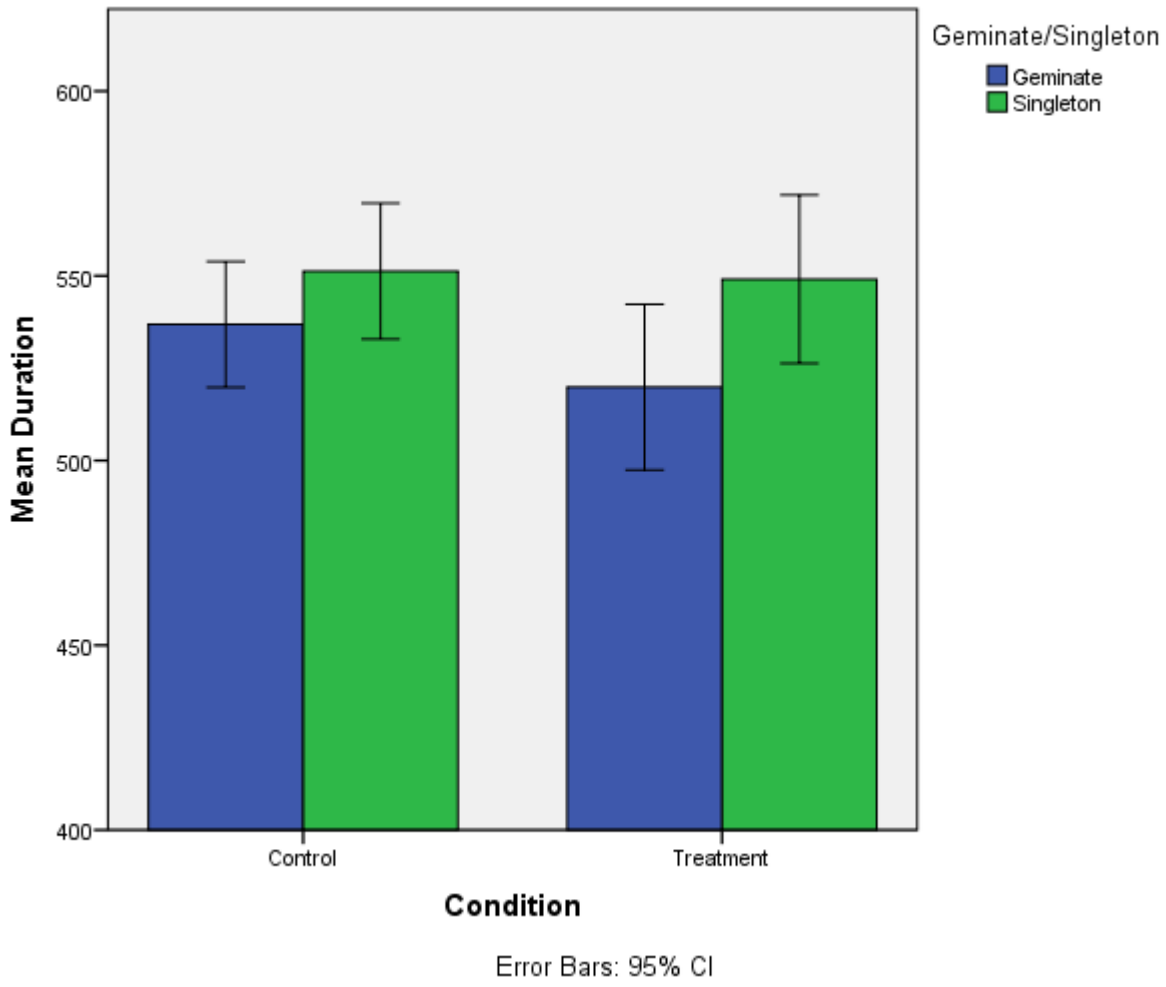


Figure 2.3: *Bar graph showing the relationship between duration and the geminate/singleton contrast in both the treatment and control conditions. Error bars indicate the 95% confidence interval.*

The graph in Figure 2.3 shows a relationship that is somewhat more complicated than that seen for f_0' in the previous section. There appears to be a relationship between duration and the geminate/singleton contrast in both the control and treatment conditions, indicating the possible

lack of an interaction based on condition, meaning that the difference between speech rate in different levels of formality could rather be a property of gemination in Japanese. However, modelling analysis reveals that the pattern is somewhat less straightforward than it might appear based on Figure 2.3. After testing for an interaction between formality and condition (with the model shown in 4a), and finding that it is *not* significant, the models in (4b) and (4c) were fitted to the data with random slopes for speaker, and sentence pair in order to test if the relationship was truly consistent and significant in both the control and treatment conditions. Speaker gender was also included as a fixed factor as there do appear to be some gender-based differences in speech rate based on the data in Table 2.4.

(4) a). ***Full Data Interaction Model***

Full Model: y = Mean Duration, β = Formality : Condition + gender,

u = Intercept + Slope of Formality | Speaker and Sentence Pair

Null Model: y = Mean Duration, β = Formality + Condition + gender,

u = Intercept + Slope of Formality | Speaker and Sentence Pair

Coefficients: FormalityF : ConditionT = 14.34 ± 13.93, t-value = 1.02

FormalityF = 14.37 ± 9.53, t-value = 1.50

ConditionT = -17.01 ± 67.45, t-value = -0.25

GenderM = -99.77 ± 13.93, t-value = -5.45

Random Effects:

Groups	Name	Variance	Std.Dev.	Corr
pair	(Intercept)	14519.4	120.50	
	FormalityF	411.3	20.28	0.10
speaker	(Intercept)	606.2	24.62	
	FormalityF	107.9	10.39	0.97
Residual		3189.5	56.48	

Model Comparison Results: $DF(X) = 1$, $X^2 = 1.0203$, $Pr(>X^2) = .3125$

b). *Treatment Model*

Full Model: y = Mean Duration, β = Formality : gender,

u = Intercept + Slope of Formality | Speaker and Sentence Pair

Null Model: y = Mean f_0, β = gender,

u = Intercept + Slope of Formality | Speaker and Sentence Pair

Coefficients: FormalityF : GenderM = 23.05 ± 14.65, t-value = 1.57

FormalityF = 19.29 ± 11.98, t-value = 1.61

GenderM = -85.8 ± 27.07, t-value = -3.17

Random Effects:

Groups	Name	Variance	Std.Dev.	Corr
speaker	(Intercept)	795.18	28.199	
	FormalityF	90.02	9.488	1.00
pair	(Intercept)	14424.21	120.101	
	FormalityF	286.00	16.912	0.14
Residual		2520.04	50.200	

Model Comparison Results: $DF(X)=2$, $X^2 = 6.4736$, $Pr(>X^2) = .039$

c). *Control Model*

Full Model: y = Mean Duration, β = Formality : gender,

u = Intercept + Slope of Formality | Speaker and Sentence Pair

Null Model: y = Mean Duration, β = gender,

u = Intercept + Slope of Formality | Speaker and Sentence Pair

Coefficients: FormalityF : GenderM = -6.73 ± 14.40, t-value = -0.46

FormalityF = 17.08 ± 11.72, t-value = 1.45

GenderM = -100.46 ± 14.40, t-value = -4.59

Random Effects:

Groups	Name	Variance	Std.Dev.	Corr
pair	(Intercept)	14599.5	120.83	
	FormalityF	490.2	22.14	0.09

speaker	(Intercept)	509.4	22.57		
	FormalityF	120.0	10.95	0.97	
Residual		3474.4			58.94

Model Comparison Results: $DF(X) = 2, X^2 = 1.9652, Pr(>X^2) = .374$

The difference in the control condition is *not* significant in spite of being visually apparent. It is, however, significant in the treatment condition. The reason for this can be seen more clearly by examining the random slopes for the random effects structure of the treatment and control models, which can be seen in Appendix 1.1. From the coefficients we can see that the pattern of shorter durations in the informal versions of the sentence pairs is less consistent than the pattern of higher mean f_0 in the informal versions; one of the speakers (speaker 3) follows the pattern but has a difference of only ~4ms in the treatment condition, and ~1ms in the control condition. Additionally, the pattern is not consistent among the different sentence pairs in either condition; in the treatment condition one of the pairs exhibits a small difference in the opposite direction from the others, while in the control condition two pairs exhibit an opposite pattern, and two others have only very small differences (~7 and ~8 ms). This greater level of variance in the control condition leads to that version of the model being non-significant. Interestingly, a univariate ANOVA *does* show the relationship between duration and the geminate/singleton contrast in the control condition to be significant, which is nicely illustrative of the danger of false positives without a full random effects structure.

Table 2.7: *Summary of modelling results in the treatment condition for the variables in this chapter. AIC = Akaike Information Criterion, an estimate of the quality of the model for the dataset. BIC = Bayesian information criterion, a measure of the likelihood of model fit. Estimate is an estimate of the overall slope of the change based on the fixed factor, with the variance caused by the random effects taken into account.*

	Model Summary					Model Comparison	
Variable	**AIC**	**BIC**	**t-value**	**Estimate**	**Std Error**	**$X^2(1)$**	**$Pr(>X^2)$**
f_0	2578.6	2626.8	-6.449	14.91 Hz	± 2.31 Hz	14.31	0.0001546
Duration	3266.8	3311.2	2.569	28.47 ms	± 11.08 ms	4.49	0.03396
Intensity	1345.7	1390.2	0.04	0.01 dB	± 0.25 dB	0.001	0.9678

2.4 Discussion

2.4.1 Intensity

Analysis of mean intensity values using mixed effects models did not show a significant relationship between intensity and the singleton/geminate contrast, and indeed appeared to show that intensity hardly varied at all. Although it is not unexpected for intensity not to correlate with the singleton/gemination (and thereby the formal/informal) contrast, it is slightly surprising that there is so little difference overall, given the varying semantic contents of the different sentences and sentence pairs. It is however possible that this consistency is due at least in part to the design of the experiment itself. Subjects were asked to speak at a consistent and natural pace, and to stay at a set distance from the microphone when speaking. In this formal experimental setting, subjects perhaps did not feel it appropriate to vary their intensity as much as they might have otherwise.

If the results are accurate however, it does suggest one interesting conclusion: that in Japanese, changes in average pitch and intensity do not appear to be correlated. Intensity has previously been shown to correlate with increased f_0 in stressed position in some languages (such as Dutch) (Ladefoged, 1971; Agaath et al, 1996), indicating that the two may be linguistically linked at some level. In reality, while the analysis did not find any significant relationship between intensity and the singleton/geminate contrast whether via ANOVA or mixed modelling analysis, an analysis of the correlation between mean f_0 and intensity using model comparison of mixed effects models *does* prove to be significant ($X^2(1) = 4.264, p < .05$). This indicates that although there is not a significant relationship between intensity and formality based on the data in this experiment, there does seem to be a positive correlation between mean f_0 and intensity. It does, however, appear that this correlation is independent of any relationship to formality, and therefore intensity will not be analyzed in the study in Chapter 3.

2.4.2 f_0

Modelling analysis showed a significant relationship between f_0 and the geminate/singleton contrast in the treatment condition, while the same relationship was not significant in the control condition. Analysis of the model coefficients fitted to the control data also does not reveal any consistent patterns, with no individual speaker or sentence pair showing change in f_0 beyond 2-3 Hz based on the singleton/geminate contrast. This supports the initial hypothesis of this study,

that the prosody of sentences containing geminate contractions would be significantly different than those that do not, lending some support to the theory that including geminate contractions in an utterance causes a shift in the speaker's speech register beyond what would be expected with only the lexical geminate/singleton contrast. There is some evidence of confounding factors in the experiment, with a significant interaction appearing between the control/treatment condition and the geminate/singleton contrast. The explanation for this interaction is not entirely clear; it is possible either that some of the subjects guessed the purpose of the experiment, and attempted to accommodate by creating a greater contrast between the pairs, or that there is some unnoticed aspect of the treatment sentences that leads to them being pronounced with a lower overall f_0. Whatever the explanation, it does indicate that there is a possible confounding variable, and that any results should be treated cautiously. In spite of this issue however, the results still appear to be robust.

As it does seem unlikely that the shift in the overall f_0 of the utterance is a correlate of the geminate/singleton contrast itself given the results in the control condition, the results of this study will be taken as evidence that there is a relationship between f_0 and formality in Japanese that bears further investigation. The results of this experiment follow the results of Ofuka and colleagues (2000) in that f_0 was found to be significantly higher in informal utterances, but contrasted with the results of Ito (2002). The corpus-based study to be described in Chapter 3

will therefore investigate the role of f_0 in the expression of formality in conversational Japanese, with the specific hypothesis that it will be raised in more informal utterances. If this hypothesis were supported, it would be interesting in that it would contradict not only Ito (2002), but also long-standing research on the prosody of politeness in Japanese (e.g. Loveday, 1981) which found that polite speech tends to be correlated with higher f_0.

2.4.3 Duration

As with the case of f_0, the durational difference was significant based on modelling analysis in the treatment condition, but not in the control condition. Although there was some evidence that durational changes in the utterance could be tied to the geminate/singleton contrast based on an apparent pattern in the control data, the analysis revealed that the pattern was only consistent in cases where the utterance contained a geminate contraction. This lends support to the hypothesis that there is a relationship between utterance duration (and due to the design of the stimuli in this experiment, possibly speech rate as well) and formality in Japanese.

Based on these results, the duration appears to be worth investigating in the corpus-based study. However, as the number of segments in the tokens of speech to be analyzed in that study cannot be as tightly controlled as in an experiment, the study will instead focus on articulation rate. This will allow for comparison of a variable between formal and informal speech without the need to be concerned with the length of the utterances.

2.5 Conclusion

The results of this study support the hypothesis that prosody plays a role in the realization of formality in Japanese, although as it used geminate contractions as a trigger to attempt to elicit informal speech from subjects in the lab rather than investigating natural informal speech the validity of the hypothesis is still somewhat uncertain. At the very least, it appears likely that the changes in prosody cannot be explained by the geminate/singleton contrast itself. While it is not possible to draw any definite conclusions from the results of this pilot study, they do indicate that the connection between formality and both f_0 and duration (and possibly articulation rate) is worth investigating further. The following chapter will describe a study which does just that, examining a corpus of conversational spoken Japanese in order to further test the hypothesis of the relationship between prosody and formality.

Chapter 3

Corpus Based Study of the Prosodic Properties of Formality in Japanese

3.1 Introduction

3.1.1 Chapter overview

This chapter describes a corpus-based study investigating the prosodic properties of the informal register of conversational speech in Japanese, and how it compares to the formal register. Japanese is a particularly appropriate language in which to study this relationship, because the large number of lexical and grammatical features indexical of speech register (Ide, 1982; Cook, 1998; Sreetharan, 2004; described further in Section 3.2) make the process of assessing the level of formality in speech less subjective than in many other languages. Based on the previous studies of the relationship between prosody and formality, and the results of the pilot study described in Chapter 2, mean f_0, articulation rate, and f_0 range were chosen as the target variables in this study. However, instead of taking a lab-based approach as in the previous socio-phonetic studies of formality in Japanese, the decision was made to create a new corpus of conversational Japanese for analysis (see Section 3.3 for further details on the decision to create a new corpus, and collection methodology), with the objective of analyzing more natural speech patterns, the formality of which could be judged post-hoc.

This chapter will first review relevant previous work regarding the relationship between prosody and formality in Japanese, and cross-linguistically. It will then give an overview of the research questions and hypotheses that this study seeks to investigate. Section 3.3 will then

describe the process of data collection and annotation used in gathering the corpus data used in the study, while Section 3.4 then describes an automated MATLAB script which was used to diagnose and correct pitch-peak estimation errors in the initial f_0 measurements of the data. Sections 3.5 and 3.6 describe the statistical analysis of the relationship between the three prosodic variables of interest and formality, using both mixed effect models, and a functional data analysis (Ramsay, 2006) respectively.

3.1.2 Background

While much work has been done examining formality (and speech register more generally) in Japanese, the majority of such work has been approached from the point of view of pragmatics (e.g. Ide, 1982; Matsumoto, 1988; Pizziconi, 2002), and largely focuses on contrasting politeness strategies in Japanese to those found in more general frameworks of politeness (such as Brown & Levinson, 1987) rather than on examining any particular phonetic aspects of speech register. These studies are of minimal relevance to this project, and thus will not be discussed in detail. There are, however, a few studies that have examined the acoustic aspects of formality in Japanese, as well as a one study of Korean whose experimental approach makes it relevant to the current study.

3.1.2.1 Studies of the prosodic properties of formality in Japanese

Previous examinations of formal/polite speech in Japanese (Loveday, 1981; Ohara, 2001;

Tsuji, 2004) have indicated that polite speech in Japanese is characterized by an increased f_0 as compared to an overall mean for both polite and non-polite speech, as predicted both by Brown & Levinson's (1987: 267-268) proposal of increased pitch as a universal indicator of increased politeness, and Ohala and Gussenhoven's frequency code (Ohala, 1984; Gussenhoven, 2002) which posits higher f_0 as a universal linguistic indicator of more formal and deferential speech. These studies largely focused on women's polite speech, and eschewed any attempt to examine informal speech alongside the formal examples, but there have been a few attempts at doing so.

Two previous studies of relevance here are Ofuka et al. (2000), an acoustic experiment regarding the prosodic cues to different levels of politeness in Japanese, and Ito (2002), a study of the effects of suprasegmentals (in this case f_0 and speech rate) on the perception of politeness in Japanese. Neither study is an exact parallel to the current one as both used elicited rather than conversational speech, but they can at least provide some initial insights into the possible prosodic properties of formality in Japanese.

Ofuka et al. (2000) involved eliciting a series of (read) utterances from speakers in a controlled laboratory setting, asking the subjects to deliver the utterances in both a 'formal' and then a 'casual' manner, and then using those recordings in a perceptual study to test the effects of different acoustic cues on a listener's perception of an utterance's politeness. The study was one of the first directly addressing the acoustic properties of casual versus formal speech, and the

target features examined were narrow – only the speech rate of the final mora and the direction of f_0 movement (i.e. rise or fall in pitch) in the final mora were investigated.

The results of the study (Ofuka et al. 2000: 213-215) were that – as foreshadowed by another study of the perception of politeness in Japanese, Ogino & Hong (1992) – both the direction of final f_0 movement and the speech rate of the final mora were used by listeners as indicators of the intended politeness of an utterance, where higher final f_0 and articulation rate indicated a more "casual" utterance. From a production standpoint the results were somewhat ambiguous – though the majority of speakers had a higher f_0 in the less polite utterances, one showed the opposite pattern. Speech rate, however, was consistently higher in the less polite utterances. These findings are relevant to the current study, in that the possibility that f_0 and speech rate are both related to an utterance's level of formality was one of the hypotheses that this study sought to test (see Section 3.2 for further discussion). However, there are some aspects of Ofuka et al. (2000) that prevented it from being more of a foundation for the current study – firstly, only two sentences were tested, and both were questions, leading to a rather semantically narrow set of stimuli. Furthermore, the fact that the recorded stimuli were all read rather than spontaneous speech, and were also collected in a lab means that the current study analyzed a very different sort of speech stimuli, which may lead to contrasting results.

Ito (2002) differs from Ofuka et al. (2000) in several ways. Firstly, it made use of a speech

corpus elicited at Chiba university (Aono et al., 1994) using the HCRC map task designed at the University of Edinburgh (Anderson et al., 1991) wherein speakers give directions to different points on a map to listeners of varying social relations to the speaker, hopefully eliciting utterances of varying levels of politeness based on the relative social standings of the conversational partners. Although these recordings were still made based on an artificial task, they should be far more natural than the read speech used in Ofuka et al. (2000). The results of Ito (2002) however were somewhat inconsistent. The study examined the overall f_0 and articulation rates of two speakers in the corpus, and these two speakers appeared to use different strategies to indicate level of formality. While one speaker appeared to increase both f_0 and articulation rate when speaking to a listener of lower status, the other speaker did not follow these patterns. Additionally, a perception experiment was conducted based on the speech of the speaker from the corpus who showed changes in f_0 and articulation rate, but listeners were found to be unable to consistently predict the level of formality based on the acoustic cues found in a single word (in this case /wakaɻimaʃita/ "I understand"). Ito (2002) did nonetheless indicate a few points of relevance for the current study. One was that it does appear possible that speakers use increased f_0 and articulation rate to indicate formality, but also that speakers may not be entirely consistent, and that in general listeners may have difficulty using isolated cues to determine formality.

3.1.2.2 Studies of the phonetic properties of formality in other languages

There are several studies of the relationship between prosody and politeness in languages other than Japanese, such as Spanish, Mandarin, Korean, and English (see e.g. Alvarez & Blondet, 2003 for Venezuelan Spanish; Lin et al., 2006 for Mandarin; Hübscher, Borràs-Comes & Prieto, 2017 for Catalan; Navarro & Nebot, 2014 for a further overview). These previous studies indicated a number of possible acoustic properties of polite speech in languages other than Japanese, such as increased f_0 height and variability in Venezuelan Spanish interrogatives (Alvarez & Blondet, 2003), and longer phrase-final durations in Mandarin (Lin et al., 2006). Of these, the studies which are most likely to be relevant to the current study are Winter & Grawunder (2012) – which investigated the relationship between acoustic factors and formality in Korean – and Hübscher et al. (2017), which was a similar study investigating Catalan Spanish. The reasons for their relevance are that, firstly, they were concerned with *formality* rather than strictly with *politeness,* which mirrored the current study. As the relationship between formality and politeness in other languages is not necessarily as clear as it is in Japanese, this is a critical point. Secondly, Winter & Grawunder (2012) appeared particularly relevant as it has been observed that there are some similarities in prosodic structure between Korean and Japanese, particularly at the level of the intonational and accentual phrases (Venditti, Jun & Beckman, 2014; Kubozono, 2015), and so making comparisons between the prosody of the two languages is not too far-fetched.

Winter & Grawunder (2012) collected speech of different levels of formality via a role-playing task, where subjects were asked to either leave a message on a cell phone voice mail, or to make a direct request of someone in person. Both scenarios were used to produce examples of formal and informal speech. The acoustic properties – including mean, range, and SD of f_0 and intensity, harmonics, articulation rate, pause count, filler count, and breath intakes – of the different levels of formality were analyzed and compared using mixed effects regression models. Significant main effects were found for mean, SD, and range of f_0 ($p < .01$ for all), articulation rate ($p < .05$), and filler count ($p < .001$), wherein all were significantly higher in informal than in formal speech. Following this 2012 study of Korean, Brown, Winter, Idemaru and Grawunder (2014) also conducted a perception experiment which tested the salience of the prosodic variables analyzed in Winter & Grawunder (2012) to Korean and English listeners' perception of the honorific speech register in Korean. While their speech stimuli were read rather than spontaneous, Brown et al. (2014) did find that both Korean and, to a lesser extent, English speaking listeners were able to use prosodic cues to correctly identify speech to a status superior or inferior at a rate greater than chance. This result implies that certain prosodic properties of different levels of politeness/formality could be present cross-linguistically.

Hübscher et al. (2017), a study of Catalan Spanish, structured similarly to Winter & Grawunder (2012) – excepting that all speech data was formulated as requests – also found significantly

higher mean f_0 in casual speech ($p < .001$). This result further contrasted the prediction of the *frequency code* (Ohala, 1984, Gussenhoven, 2002) which states that higher f_0 should correlate with formal and deferential speech. Hübscher et al. (2017) also found that there was a higher pause rate, lower intensity, and slower speech rate in formal speech, which they termed 'prosodic mitigation'. The similarity of these results to Winter & Grawunder (2012) indicated that this strategy of prosodic mitigation is used in multiple languages in the expression of formality, and that the relevant variables were worth investigating in the current study.

Although the results of Winter & Grawunder (2012) and Brown et al. (2014) were not necessarily enough by themselves to hypothesize that the same results would be seen in the current study, the observed similarity between the prosody of the two languages allowed them to serve as a point of reference. The similar results of Hübscher et al (2017) do suggest some possible cross-linguistic patterns in the relationship between prosody and formality in speech, and the fact that many aspects of f_0 (mean, SD, and range) and intensity (SD and range) appeared to co-vary significantly with changes in formality in both Korean and Catalan indicated that it was worth investigating the relationship between these variables and formality in Japanese to a similar (or greater) level of depth. Additionally, the fact that many of the tested phonetic parameters were higher in informal speech was of interest, as this result could help better inform the hypotheses of the current study as they related to the expected relationship between prosody and formality.

Comparing the results of this study of Japanese to Winter & Grawunder (2012) and Hübscher et al (2017) can also provide some insight as to the possible presence of some of these phonetic cues to formality across multiple languages, as if the results are very similar it could indicate that there are cross-linguistic tendencies in how prosody relates to speech register.

3.2 Research questions and hypotheses

3.2.1 Research questions

As was discussed in Chapter 1, this study investigates two related research questions: 1) *Do speakers of Japanese make use of changes in prosody to help communicate their intended level of formality in conversation?* and if that is the case 2) *what* specific *differences in the prosody of an utterance are used by speakers to distinguish different levels of formality?* These questions have been partially addressed by the pilot study in Chapter 2 which found a significant relationship of mean f_0 and utterance duration to prosody wherein mean f_0 was higher and mean duration lower (indicating faster speech) in informal utterances. However, the pilot study cannot definitively answer either of these questions – although the results are not necessarily incorrect, the small scale of the pilot, along with the facts that the speech analyzed was read, and that different levels of formality were elicited indirectly casts some doubt on the pilot study's conclusions. The current study will investigate questions (1) and (2) using a corpus of natural

conversational Japanese speech, and the formality of each utterance will be judged based on a consistent set of criteria, discussed further in Section 3.3.2.

3.2.2 Hypotheses

The results of the pilot study (and to a lesser extent the results from the previous literature) have led to two hypotheses: 1) *Speakers do use changes in prosody to express different levels of formality in speech,* and 2) *Differences in mean f_0, articulation rate, and f_0 range will co-vary significantly with the level of formality in speech. Each of these variables will be higher in informal speech.* Hypothesis (1) has already seen some support from the results of the pilot study, although as discussed in Section 3.2.1 the results are not conclusive. This study will test the hypothesized relationship between prosody and formality on a much larger scale, and in more natural speech, hopefully leading to more robust results. Hypothesis (2) is based largely on the facts that mean f_0 was significantly higher and duration significantly shorter (indicating higher articulation rate) in informal utterances in the pilot study. The part of the hypothesis predicting that f_0 range will also be significantly higher in informal speech, however, is based on the findings from Winter & Grawunder (2012) which showed that f_0 range was significantly higher in informal speech in Korean. Observational evidence from the pilot study data suggests that this could be the case in Japanese as well, and so the decision was made to test f_0 range in addition to the variables tested in the pilot study.

3.3 Data collection and annotation

As the goal of the experiment is primarily to examine the properties of conversational Japanese, the decision was made to conduct a corpus-based experiment. Although any speech recorded outside of natural conversation is likely to fall somewhat short of the ideal level of naturalness for this study, spontaneous speech gathered for a corpus is likely to come much closer to the desired speech registers than speech obtained in a lab. Additionally, speech corpus data has in the past produced significant results in phonetic studies where similar lab based experiments did not (such as Gahl, 2008 vs. Guion, 1995).

Initial investigations into possible corpora of spoken Japanese revealed two potentially viable sources: The Corpus of Spontaneous Japanese (Maekawa, 2003), and the Chiba University 3-way conversation corpus (Den, 2014). Both corpora were gathered in Japan, and contain entirely spontaneous speech. The CSJ contains mainly monologues, and presentations, while the Chiba corpus contains short conversations between 3 participants. Both corpora are very well annotated, with detailed segment meta-data and pitch tracking information, but both also have shortcomings.

The main shortcoming of the CSJ for this study is that it although it contains millions of words, it contains very little speech that could be considered informal, and essentially none that could be considered conversational. The Chiba corpus on the other hand contains entirely conversational speech of varying levels of formality, but unfortunately only consists of roughly

30 minutes of recordings, which would not be sufficient for the scope of this study. Because of the shortcomings of the available corpora, and in order to have greater control over the collection methods and content of the data, the decision was made to create a new small corpus of conversational Japanese speech.

3.3.1 Data collection methodology

The speech data for this study was collected at the NINJAL institute in Tachikawa-shi, Japan via one-on-one interviews between the experimenter and subjects. The interviewer was a non-native speaker of Japanese with a high-level of proficiency, and subjects were 10 native speakers of Japanese aged 31-45 (5 male, 5 female). The age of subjects was kept below 50 in order to minimize any potential influence of the effects of age on f_0 (Harrington et al., 2007), and all subjects were speakers of the Tokyo dialect of Japanese (born and raised in the Tokyo area up to age 18) in order to reduce any possible effects on f_0 from different dialects (Kubozono, 2012). Interviews were conducted in a lounge setting rather than a recording booth or lab in order to encourage a more natural, conversational style of speech. Although the topics discussed in the interviews were not completely consistent, certain topics were brought up frequently and are listed in Appendix III. Recordings were single-channel mono, made at 48 kHz, 16-bit PCM.

The format of the interviews was similar to a sociolinguistic interview (Labov, 1972) but with less control of the topics discussed, and each subject was recorded for ~30 minutes. All

interviews began with self-introductions from both the interviewer and the subject, which were generally quite formal, and then proceeded naturally to other topics as they arose, with the interviewer gradually modulating their speech register to a more informal level to encourage the subject to follow. In general, this resulted in a pattern where the first five minutes of the interview consisted mainly of formal speech, minutes 5-10 consisted of a mix of formal and informal speech, with subjects sometimes code-switching within utterances, and the remainder of the interview consisting of mostly informal speech. At the end of each interview, subjects were asked to read a short passage to provide an example of read speech, to use as a control.

In total this resulted in ~5 hours of recorded speech for analysis.

3.3.2 Data annotation

To ensure analysis of the correct portions of the recorded data, it was necessary to carefully segment and annotate the extended recordings. As the target of the experiment is only utterances by the subjects, these targets had to be separated from the interviewer's speech. In this case (and in the remainder of this chapter) the term “Utterances” refers specifically to portions of speech which have been delimited in the following way. Portions of the subjects’ speech were manually labelled as intervals in a Praat text grid. Interval labels were created for each instance of speech by the subjects, with the following exceptions:

- Isolated filler interjections (such as /e:/ or /a:/) were not included.

- Isolated laughter was not included, unless it occurred clause-internally.
- Extended pauses (defined as pauses of >1 second) were not included.

Boundary labels were placed either at clause boundaries for full sentences, at the start/end of an extended pause for fragments, or at turn-taking boundaries (the boundary between the end of the first speaker's turn and the start of the second speaker's) in the case of back-and-forth conversation containing fragments.

The number of lexical moras within each clause or fragment was manually counted in order to allow the calculation of articulation rate data. Pauses of less than 1s were included and counted as 1 mora per 100 ms consistently in order to reduce any effects of more or less frequent pausing in speech on articulation rate. Although previous study of spontaneous speech in Japanese has shown that moras are not uniform in duration (Warner & Arai, 2000), their counting was kept consistent for the sake of keeping the effect of more or less pausing consistent between speakers. An example of a partial waveform labelled in a Praat TextGrid is shown in Figure 3.1.

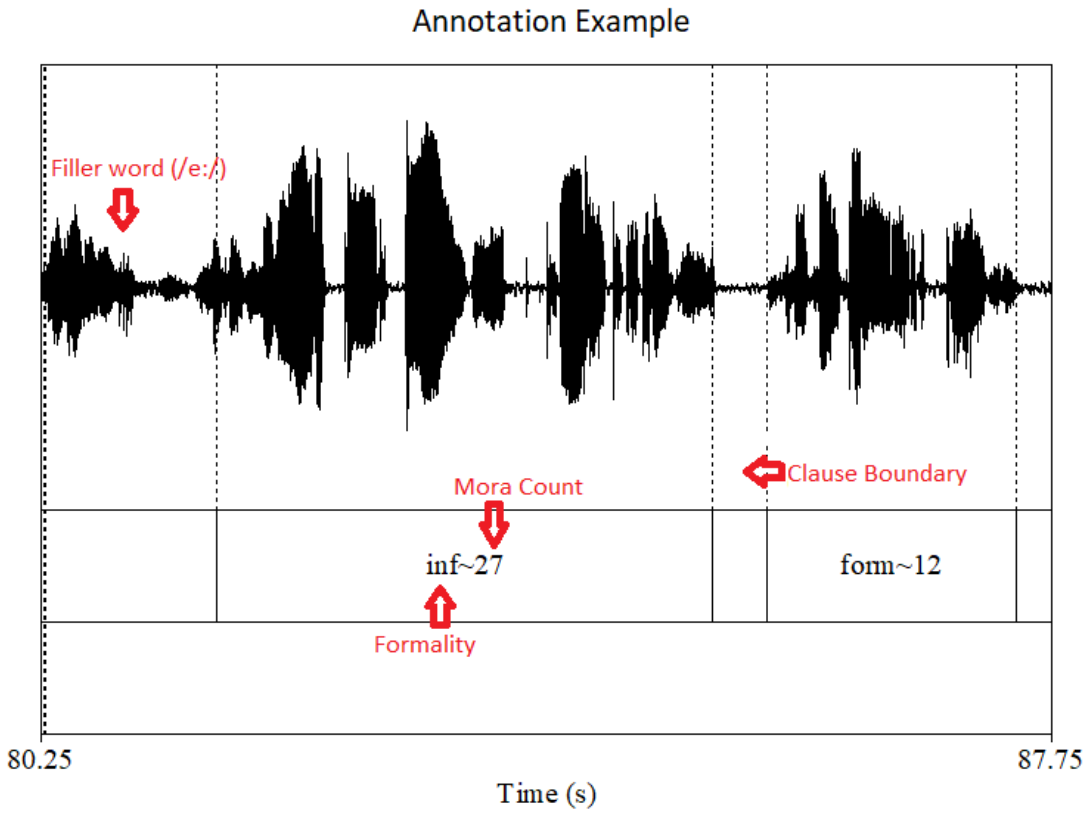


Figure 3.1: *An example of a waveform annotated in a Praat TextGrid. In this particular portion of the waveform, one utterance was judged to be informal, and one was formal. An initial filler word (/e:/) was not included, and boundary labels were placed at a clause boundary.*

The speech within each labeled interval was judged to be either formal, informal, or read. This judgment was initially made by the experimenter, and once each interval was assessed, linguistically naive speakers of Japanese were later asked to judge the formality of randomly selected intervals in order to confirm the judgments. In any cases where there were differences in judgments – which occurred in roughly 2.5% of intervals – the assessment of the native speaker was followed. Although any judgment of the level of formality of a given utterance will be to

some degree in the eye of the beholder, because the determination of levels of formality is very important to this study a consistent set of criteria to judge formality was established (Table 3.1).

Table 3.1: *Criteria used in determining utterance formality.*

Criteria	Formal Example	Informal Example
Copular verb	/desɯ̟/ "to be"	/da/ "to be"
Verb form	/ʃimaʃita/ "did"	/ʃita/ "did"
Sentence-final particles	/-wa/	/-jo/
Question particles	/-ka/	/-kai/
Under/over articulation	/tsɯ̟maɻanai/ "boring"	/tsɯ̟man:ai/ "boring"
Indexical word forms	/jahaɻi/ "...after all"	/jap:a/ "...after all"
Honorifics	/ikaɻemasɯ̟/ "to go (HON)"	/ikɯ̟/ "to go"

The criteria in Table 3.1 were determined largely by previous examinations of lexical items and phonological forms indexical of different registers of formality in Japanese (Ide, 1982; Cook, 1998; Okamoto, 1999) and of observational evidence of spoken Japanese. Although these criteria were applied consistently, there was a small percentage of utterances which were ambiguous either due to a lack of criteria present in the utterance, or code switching. An example of one such case is given in (1).

(1) Kyonen kankoku-ni it-ta-n desu yo. ("I went to Korea last year")

Last year Korea-DAT go-PAST-NOM Copula Intensifier

The sentence in (1) was initially judged to be formal due to the presence of the formal copula "desu". However, it was judged by a native speaker to be *informal*, possibly due to the presence

of the intensifier "yo", and the fact that the speaker used a verb construction that is somewhat emotive (the inclusion of the nominalizing "n" after the standard past tense).

Once all the speech data was labeled, it was then automatically segmented into separate .wav files (one per labeled interval), and f_0 and articulation rate data was measured using Praat and bash scripts. In total, this resulted in 2,697 utterances, of which 416 were formal, 2,067 were informal, and 214 were read. There were 1,314 utterances by female subjects and 1,383 by male subjects.

3.4 f_0 measurement and correction

As the measurement and analysis of mean f_0 and f_0 range are central to this study, it is important to be certain that the f_0 measurements are as accurate as possible. However, the initial measurements from Praat are not entirely reliable, due to pitch peak estimation errors (i.e. pitch doubling) (Kochanski, 2010), as can be seen in Figure 3.2.

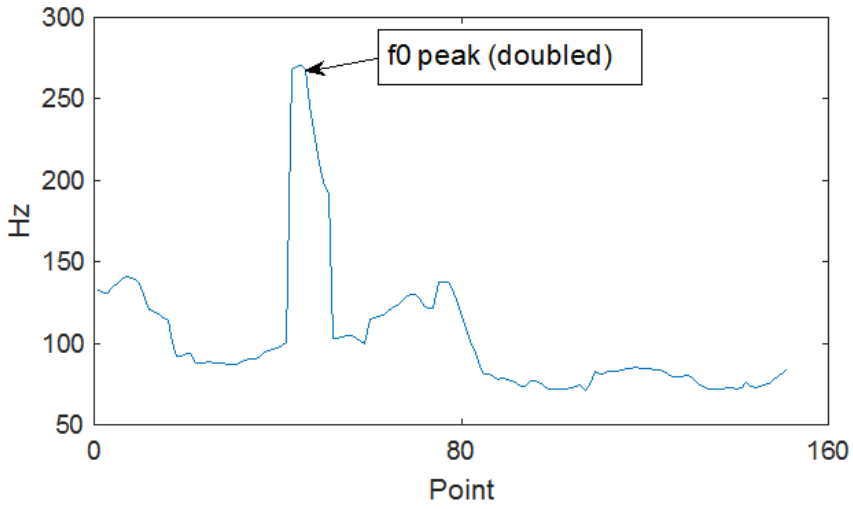


Figure 3.2: *A f_0 vector showing a pitch-doubling error.*

Such errors present a large problem for the analysis of both mean f_0 and f_0 range, as although it is possible that the errors are spread proportionally among the different levels of formality, it is not possible to determine this without examining every single utterance. It is therefore difficult to know to what degree pitch tracking errors impact the analysis, and because not all utterances contain such errors simply halving pitch peaks does not appear to be a viable solution. To help overcome this problem, an automated MATLAB script to diagnose and fix pitch-doubling errors in f_0 vectors was developed.

3.4.1 Pitch doubling errors

In order to accurately analyze f_0, it is important to find a method to correct as many pitch-doubling errors in the f_0 vectors as possible. As can be seen in Figures 3.2 and 3.3, these errors are typically

somewhat square-shaped, and tend to result in a very sudden increase in the slope of the vector. Based on this, it appears that the simplest way to diagnose these pitch peak errors is to examine the first differences (the differences between each pair of consecutive points) of the vector, and search for changes that fall outside of an expected range of tolerance. After calculating the first differences and testing a MATLAB script which attempted to diagnose pitch-tracking errors and testing it on a number of f_0 vectors, a threshold of three times the mean of the absolute value (i.e. with positive and negative differences treated the same) of the first differences was arrived at.

After testing for errors, the vector was then corrected by adjusting the values between two changes in the slope of the vector outside of the set level of tolerance by the difference between the f_0 value at the peak of the error and the point immediately after the end of the error. There are a few caveats to this: the script also corrects errors in the vector where the pitch appears to be halved (i.e. trough estimation errors) as it tests the absolute value of the first differences against the tolerance. The script will also avoid correcting extremely long errors (defined as lasting for more than 15% of the total length of the vector; this is because most of these sudden drops or increases in f_0 followed by an extended portion of the vector that appears normal generally result from the fact that undefined f_0 values given by the Praat pitch tracking are ignored, and therefore any gaps in voicing are not represented in the vector. This leads to there being apparently sudden

changes in the f_0 vector which are actually correct. Figure 3.3 shows an example of a f_0 vector which has been corrected by the MATLAB script.

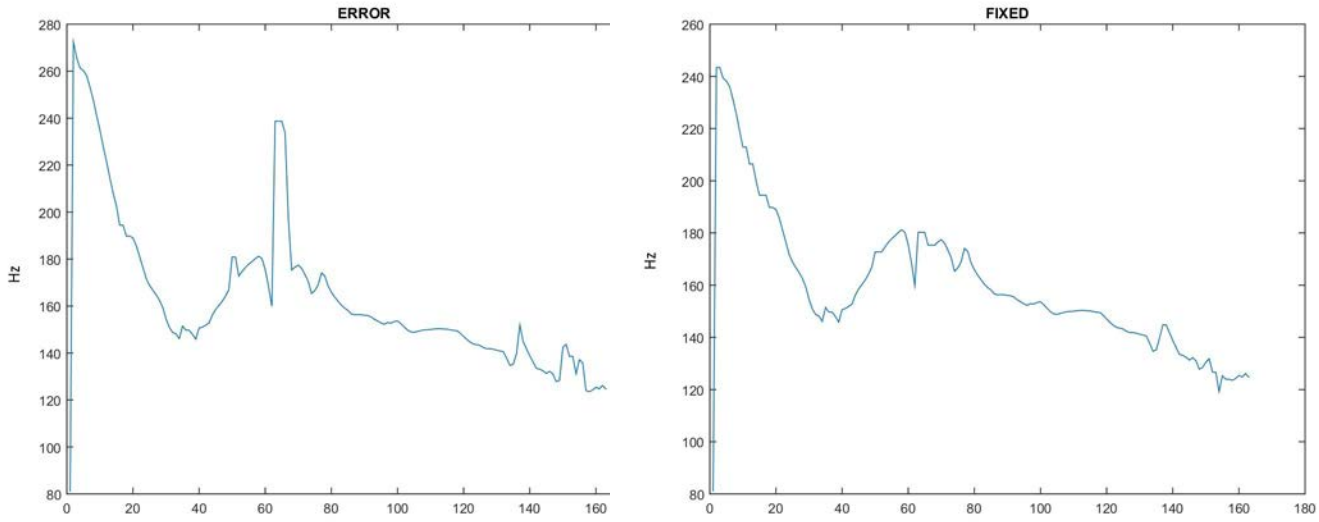


Figure 3.3: *f_0 vectors with a pitch-doubling error (on the left) and after automatic correction (on the right).*

With those fixed f_0 vectors created, it is now possible to move on to the next step of analyzing the f_0 measurements.

3.5 Data analysis

The variables analyzed in this study were mean f_0, articulation rate, and f_0 range (defined as four standard deviations of f_0). f_0 was measured using Praat scripts. As discussed in Section 3.2.2, the hypothesis the study seeks to test is that, based on the pilot study, and on a similar study of the acoustic properties of formality in Korean (Winter & Grawunder, 2012) each variable will be

significantly higher in informal speech than in formal speech.

3.5.1 Articulation rate

Table 3.2 shows articulation rate statistics for each level of formality.

Table 3.2: *Articulation rate statistics.*

Formality	Mean	Std. Deviation
Informal	7.82 moras/second	1.51 m/s (19% of mean)
Formal	6.64 moras/second	1.63 m/s (24% of mean)
Read	6.93 moras/second	1.16 m/s (17% of mean)

It is immediately apparent from Table 3.2 that the mean articulation rate of informal speech is quite a bit higher than that of either formal or read. Model comparison using a likelihood-ratio test of a linear mixed-effects model in R, shown in (2), shows a significant relationship between articulation rate and formality:

(2) Full Model: y = Mean Rate, β = Formality : gender,

u = Intercept + Slope of Formality | Speaker

Null Model: y = Mean Rate, β = gender,

u = Intercept + Slope of Formality | Speaker

Coefficients: FormalityF : GenderM = -0.20 ± 0.21, t-value = -0.97

FormalityF = -1.12 ± 0.14, t-value = -7.53

GenderM = -0.17 ± 0.47, t-value = -0.37

Random Effects:

Groups	Name	Variance	Std.Dev.	Corr
speaker	(Intercept)	0.54970	0.7414	
	formality	0.05674	0.2382	-0.73
Residual		1.77078	1.3307	

Model Comparison Results: Formality: $DF(X) = 2, X^2 = 26.96, Pr(>X^2) = < .001$

Gender: $DF(X) = 2, X^2 = 1.98, Pr(>X^2) = .37$

Interaction: $DF(X) = 1, X^2 = 0.92, Pr(>X^2) = .33$

This initial finding agrees with the results of the pilot study, and of previous acoustic studies of Japanese informal speech (Ofuka et al., 2000; Ito, 2002), but does not tell the whole story of the relationship between articulation rate and formality. Figure 3.4 compares histograms of articulation rate in formal and informal speech.

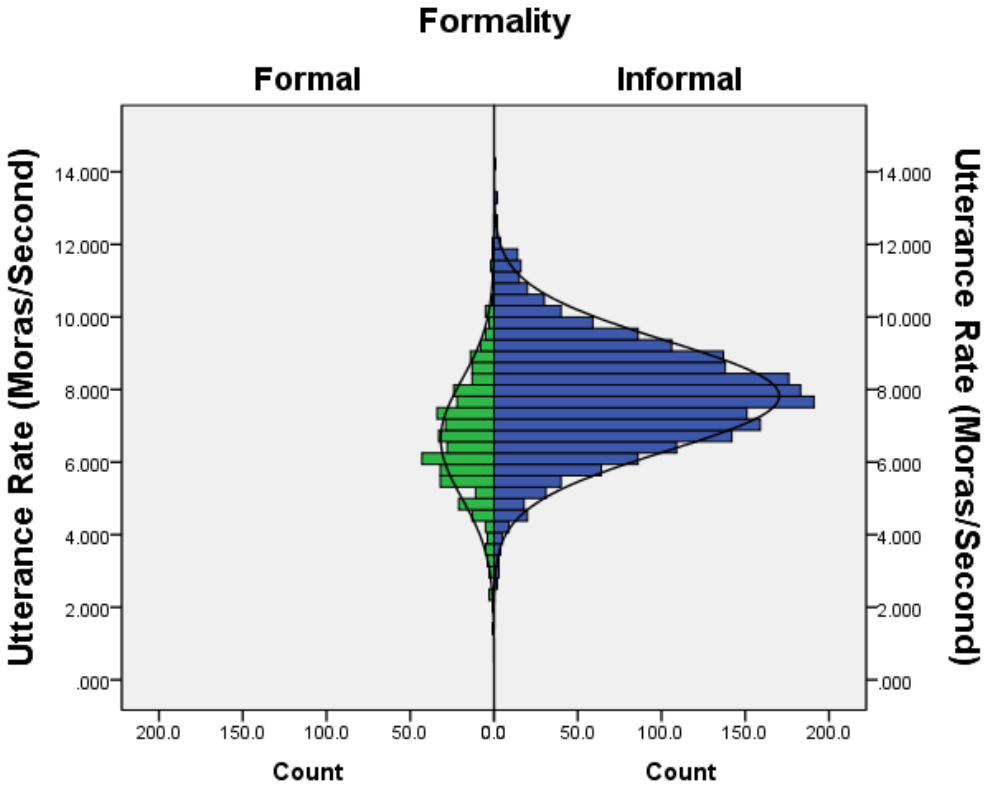


Figure 3.4: *Histogram of articulation rate in informal and formal speech.*

It is apparent from the distributions in Figure 3.4 that although both informal and formal speech have a similar minimum for articulation rate (~2 moras/second), informal speech appears to have a much higher maximum articulation rate (topping out at around 12 moras/second, with a few

outliers that are even faster). This indicates that although speakers did not always articulate faster in informal speech (although they did so typically), there was a greater possible range of articulation rates in informal speech.

These differences in articulation rate hold regardless of other factors, with linear mixed effects regressions showing no significant change with the addition of random effects of speaker age, or gender.

3.5.2 Mean f_0

Mean f_0 is also different in formal and informal speech. Table 3.3 shows the f_0 data for each level of formality.

Table 3.3: *Mean f_0 statistics.*

Formality	Mean f_0	Std. Deviation
Informal	166.4 Hz	49.8 Hz (29% of mean)
Formal	151.0 Hz	45.6 Hz (30% of mean)
Read	155.4 Hz	40.9 Hz (26% of mean)

Based on the values in Table 3.3 f_0 also appeared to be a potentially significant cue to the formal vs. informal contrast, but there were some issues with the data that are important to note. Figure 3.5 shows a density plot of mean f_0 for each gender.

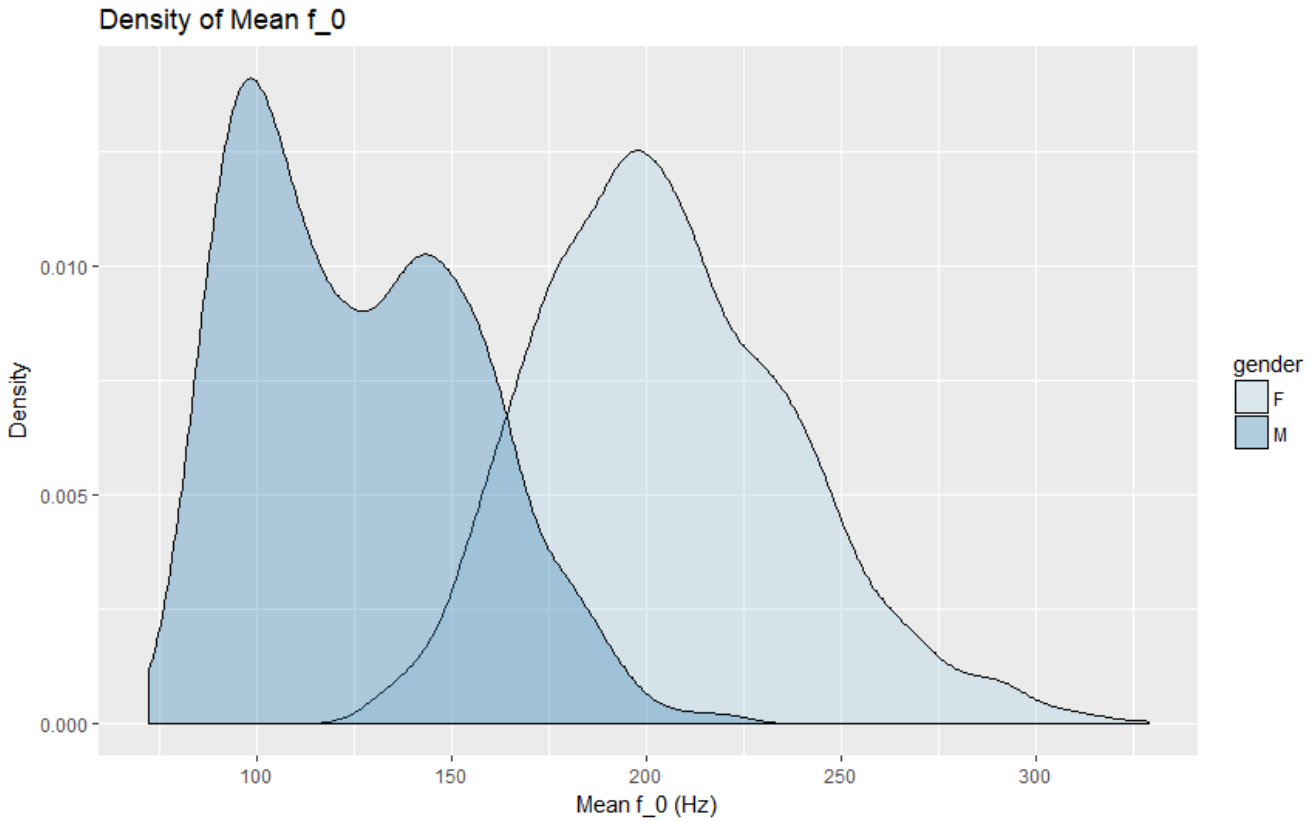


Figure 3.5: *Density plot of mean f_0 for male and female speakers.*

It is readily apparent that f_0 is not normally distributed, particularly for male speakers, who appear to have two separate peaks of density. Although it is possible that this data would still be interpretable in a linear mixed effects model with the inclusion of sufficient random factors to normalize the residuals (Bates et al., 2015), it is in general better statistical practice to either normalize the data, or (in the case of binary data as in the current study) make use of generalized linear mixed effects models (see Bolker et al., 2009 for some discussion). Since it is already generally acceptable to present f_0 on a non-linear scale (Stevens & Volkman, 1940; Traunmüller,

1981; Fujisaki & Hirose, 1984; Henton, 1989; Nolan, 2003), the mean f_0 values were $\log_{10}$ transformed for analysis.

(3) **Full Model**: y = $\log_{10}$ Mean f_0, β = Formality : gender,

u = Intercept + Slope of Formality | Speaker

Null Model: y = $\log_{10}$ Mean f_0, β = gender,

u = Intercept + Slope of Formality | Speaker

Coefficients: FormalityF : GenderM = -0.002 ± 0.006, t-value = -0.38

FormalityF = -0.038 ± 0.004, t-value = -8.391

GenderM = -0.221 ± 0.039, t-value = -5.59

Random Effects:

Groups	**Name**	**Variance**	**Std.Dev.**	**Corr**
speaker	(Intercept)	3.945e-03	0.062811	
	formality	6.899e-06	0.002627	1.00
Residual		3.221e-03	0.056753	

Model Comparison Results: Formality: $DF(X) = 2, X^2 = 30.48, Pr(>X^2) = < .001$

Gender: $DF(X) = 2, X^2 = 14.33, Pr(>X^2) = < .001$

Interaction: $DF(X) = 1, X^2 = 0.14, Pr(>X^2) = .70$

Additionally, mean f_0 and articulation rate do not appear to be confounded, with both a linear mixed effects regression and a Pearson correlation showing no significant relationship, meaning that the mean increase in f_0 does not appear to be caused by the overall mean increase in articulation rate in informal speech.

3.5.3 f_0 range

The final variable tested was f_0 range, here defined as the difference between +/- two standard deviations of f_0 (e.g. 4 standard deviations). Although this is a somewhat simple measure, it can still be of use for determining general patterns, and as a single variable representative of a speaker's pitch range. Table 3.4 shows the f_0 range statistics for each level of formality.

Table 3.4*: f_0 range statistics.*

Formality	Mean f_0 range	Std. Deviation
Informal	131.09 Hz	49.54 Hz (37% of mean)
Formal	94.81 Hz	50.90 Hz (53% of mean)
Read	105.35 Hz	58.45 Hz (55% of mean)

It is immediately apparent when observing the descriptives in Table 3.4 that the values for f_0 range are markedly different between formal and informal speech. Although the standard deviations of these measurements are relatively high, indicating a large amount of variability in the possible ranges, they are somewhat less so for informal speech, meaning that a wider pitch range may be a more inflexible property of informal speech than in other registers. Model comparison of a linear mixed effects regression shown in (4) indicates that there is a significant relationship between formality and f_0 range.

(4) Full Model: $y =$ Mean Range, $\beta =$ Formality : gender,

$u =$ Intercept + Slope of Formality | Speaker

Null Model: $y =$ Mean Range, $\beta =$ gender,

u = Intercept + Slope of Formality | Speaker

Coefficients: FormalityF : GenderM = 7.06 ± 6.14, t-value = -0.38

FormalityF = -39.58 ± 4.41, t-value = -8.96

GenderM = -36.09 ± 13.02, t-value = -2.77

Random Effects:

Groups	Name	Variance	Std.Dev.	Corr
speaker	(Intercept)	411.56	20.287	
	formality	18.85	4.342	-0.18
Residual		2571.93	50.714	

Model Comparison Results: Formality: $DF(X) = 2, X^2 = 27.49, Pr(>X^2) = < .001$

Gender: $DF(X) = 2, X^2 = 6.05, Pr(>X^2) = < .05$

Interaction: $DF(X) = 1, X^2 = 1.26, Pr(>X^2) = .26$

Speaker gender is included as a fixed factor because f_0 range has previously been shown to have a significant relationship with gender (Ohara, 2004), with female speakers having a significantly larger range than males. The significance of this model, in combination with the others described in Section 3.5, shows strong evidence for a relationship between a number of prosodic factors and formality in speech. A summary of the results of all the modeling analyses in this section can be seen in Table 3.5.

Table 3.5: *Summary of modeling results for the variables in this chapter. AIC = Akaike Information Criterion, an estimate of the quality of the model for the data set. BIC = Bayesian information criterion, a measure of the likelihood of model fit. Estimate is an estimate of the overall slope of the change based on the fixed factor, with the variance caused by the random effects taken into account.*

	Model Summary					Model Comparison	
Variable	**AIC**	**BIC**	**t-value**	**Estimate**	**Std Error**	**$X^2(2)$**	**$Pr(>X^2)$**
Mean f_0	22512.6	22570.7	-10.066	-14.182	± 1.409	24.451	< .001
Articulation Rate	8526.4	8561.3	-11.34	-1.2285	± 0.1083	26.092	< .001
f_0 Range	26604.1	26662.2	-11.038	-35.875	± 3.250	26.188	< .001

However, although the high standard deviation of the values in Table 3.4 do not invalidate the analysis of f_0 range, they do at the very least demand a more in-depth analysis of the variable. In order to accomplish this, as well as to provide a more detailed analysis of f_0 as a whole, a *functional data analysis* was adopted.

3.6 Functional data analysis

Functional data analysis refers to a methodology whereby continuous functions (in this case orthogonal polynomials) are fitted to discretely sampled data, and the coefficients of the fitted polynomials are related to linguistic variables (Grabe et al., 2007; Ramsay, 2006). This was done by using the *polyfit* function in *MATLAB* in order to initially fit cubic functions to f_0 vectors.

First, the pitch tracking data which had been fixed by the script described in Section 3.4 was taken, and the vectors were normalized using the operation in (5), and then normalized for

time (where *y* is the original f_0 vector, and *yn* is the normalized vector, centered on 0).

(5) $$yn = \frac{y}{mean(y)-1}$$

In order to further determine the goodness of fit, the sum of the squared differences between the fitted function and the normalized data vector was calculated using the operation in (6) (where *yf* is the fitted function).

(6) $$d = \frac{\sum(yf-yn)^2}{length(yn)}$$

However, an examination of how the resulting functions fitted some of the longer utterances revealed a major problem with this initial approach. When a cubic function was fitted to each utterance, the mean *d* – as in (6) – was .097. For reference, Figure 3.6 shows an example of a fitted cubic function for a random utterance from the data with a *d* equaling approximately 0.1.

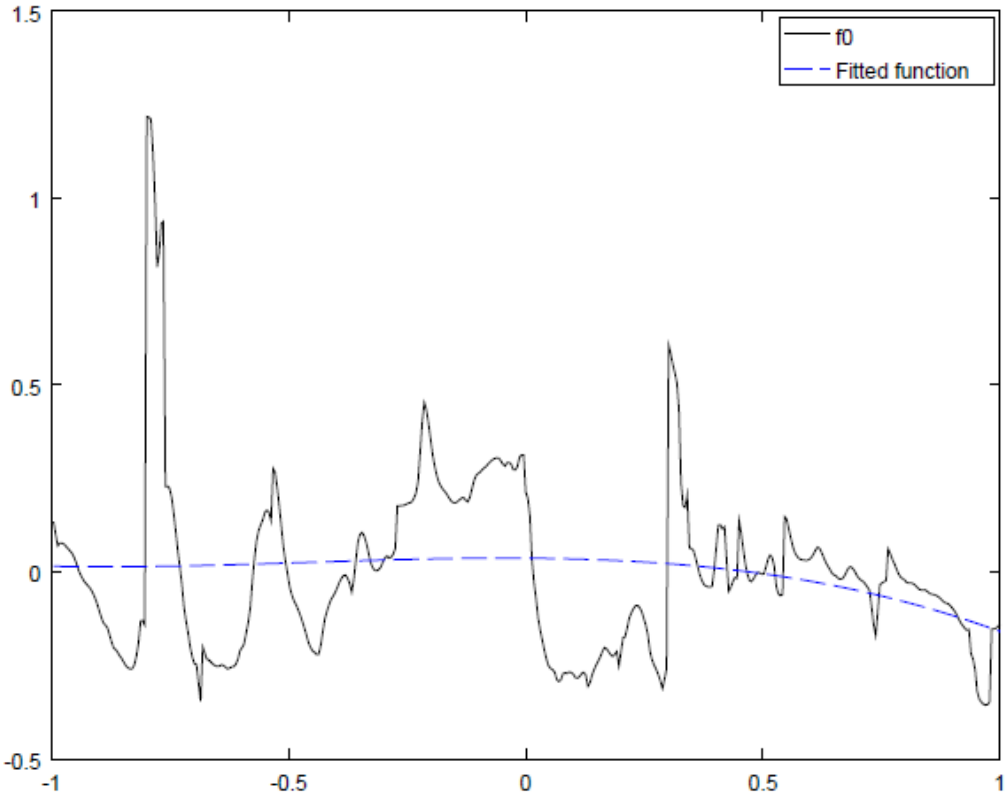


Figure 3.6: *Example of a fitted function with a d of 0.1.*

As can be seen clearly from Figure 3.6, with a d of 0.1, the fitted function is essentially meaningless, fitting to almost no part of the original vector. An average difference of this magnitude means that there is little chance of obtaining any meaningful data.

Closer examination of the fitted functions compared to the f_0 contours made it apparent that a d of around .02 - .04 was ideal for the function to fit accurately and smooth out some of the jagged movement of the vector not fixed by the script described in Section 3.4. Figure 3.7 shows an example of such a function.

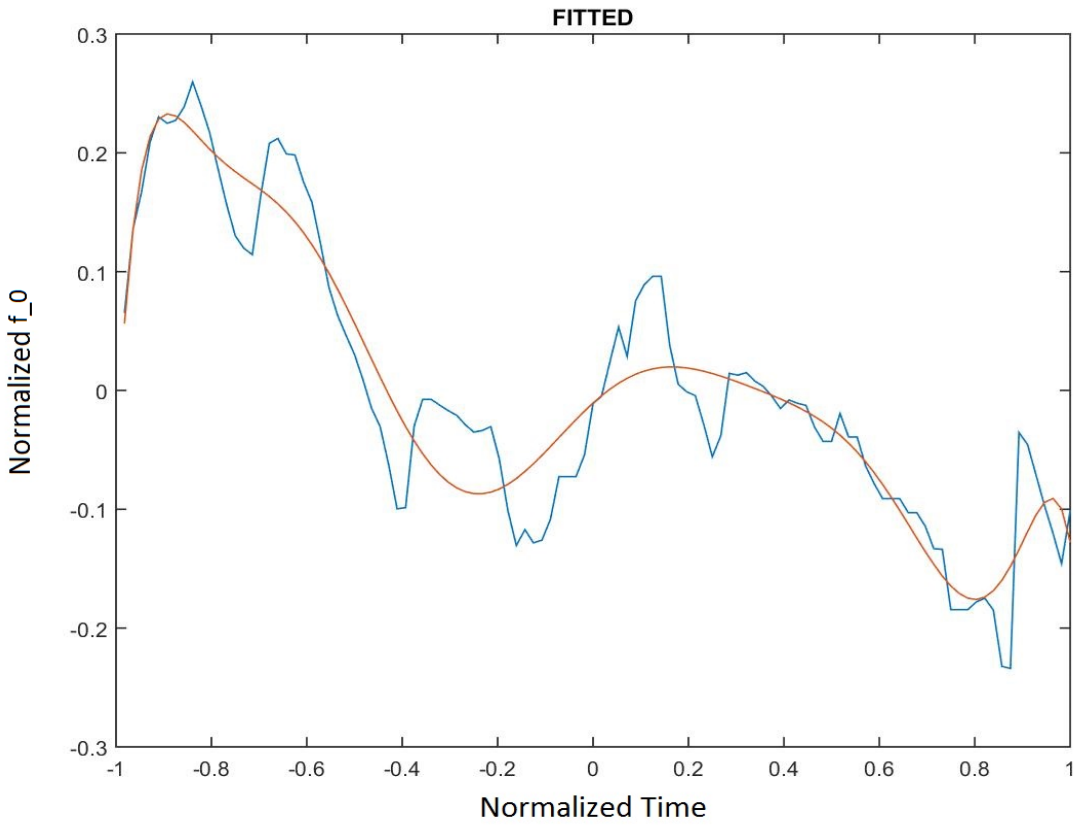


Figure 3.7: *Fitted function with a d of ~.02.*

The function in Figure 3.7 fits the vector reasonably well, and also appears to smooth out some small, rapid jumps and drops in f_0 that would not be detected as errors by the script in Section 3.4. In order to achieve a d of 0.02 - 0.04 for all utterances, another function was written in MATLAB which began by attempting to fit a quintic function to each vector, tested the goodness of fit of the function after fitting, and then either increased the degree (up to a 25 degree polynomial, beyond which little goodness of fit is gained) if it was not a good enough fit, or decreased the degree if it was possible to lose some coefficients while maintaining accuracy (down to a cubic polynomial at lowest). However, with such a large number of possible

coefficients it is difficult to relate each one to a linguistic variable, so in addition to taking the longer fitted functions a second method was adopted.

The fitted functions were each broken down into trough-to-trough sections to be analyzed. In other words, after fitting, the fitted function was divided into the following segments:

- From the start of the vector to the first trough.
- The interval between each trough and the next.
- From the final trough to the end of the vector.

In terms of the relationship of these trough-to-trough sections to the actual prosody, each section analyzed was roughly equivalent to one *accentual phrase* in Japanese (Beckman & Pierrehumbert, 1986; Pierrehumbert & Beckman, 1988; Kubozono, 1993). An accentual phrase in (Tokyo) Japanese is characterized by an initial lowering of the pitch, followed by a rise in pitch up to a 'pitch accent' (see Kubozono, 2011 for a summary of the Japanese pitch accent system), after which there may be a fall in pitch dependent upon the location of the pitch-accent (Pierrehumbert & Beckman, 1988). Since the point at which pitch begins rising generally marks the beginning of the accentual phrase, the functions analyzed – which start at pitch troughs – should correspond approximately to the accentual phrase structure of the utterance.

Once the portions of the initial fitted function were broken down, cubic polynomials were then fitted to each of those portions of the original (normalized) f_0 contour that matched the

extracted portions of the fitted function, and the orthogonalized coefficients of these new cubic polynomials were analyzed. This resulted in, on average, an excellent function fit, with a mean *d* of the cubic functions obtained using this method of 0.019. The four orthogonalized coefficients of those functions can be interpreted as follows (Grabe et al., 2007):

1. Coefficient 1 maps to the S-shaped 'wiggle' of the function, i.e. how much it moves up and down. Given that this method analyzed from trough-to-trough, this coefficient is unlikely to be high.
2. Coefficient 2 corresponds to the breadth of curvature of the function, or how sharply the f_0 rises towards and falls from the peak. This is broadly equivalent to *pitch dynamism* (Henton, 1989).
3. Coefficient 3 corresponds to the slope of the function, or how steeply the f_0 rises or falls overall. This relates to the height of the peak of the vector.
4. Coefficient 4 (the intercept) corresponds to the average height of the function, (i.e. the mean f_0)

In total, this resulted in 28,841 sets of coefficients. To avoid skewing the data with poorly fitted functions, any function with a *d* greater than 0.04 was removed from the data set, resulting in 26,493 total coefficient sets to be analyzed. A comparison of the average peak shape of the average functions (obtained by taking the mean of each of the four coefficients for informal and

formal speech) can be seen in Figure 3.8, and a list of the mean orthogonalized coefficients can be seen in Table 3.6. The average peak shapes seen in Figure 3.8 should be approximately equivalent to the average f_0 contour of an accentual phrase in each level of formality.

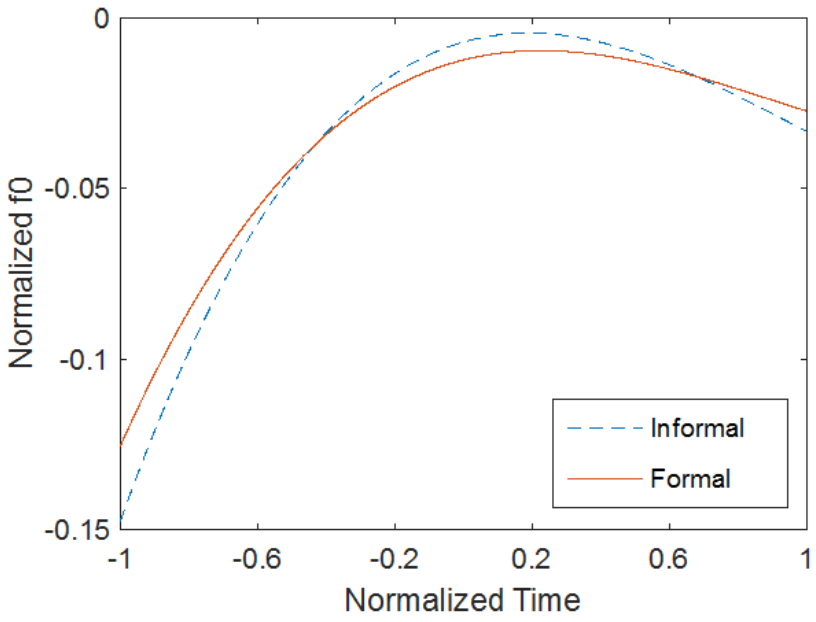


Figure 3.8: *Average peak shape of the fitted functions for informal and formal speech.*

Table 3.6: *List of mean orthogonalized coefficients.*

Formality	Coeff. 1	Coeff. 2*	Coeff. 3*	Coeff. 4*
Informal	.0285	-.0834	.0287	-.0072
Formal	.0237	-.0641	.0252	-.0122

** This coefficient is significantly different in a generalized linear mixed-effects regression model.*

There are a few visually apparent differences between the average peaks for formal and informal speech. The function representing an accentual phrase in informal speech appears to start lower, and peak higher (related to coefficient 3), while also rising and falling more sharply (coefficient 2). The shape of this function indicates both that the initial pitch lowering observed in accentual

phrases in Tokyo Japanese (Pierrehumbert & Beckman, 1988) is more pronounced in informal speech, and that informal accentual phrases appear to have greater pitch dynamism (Henton, 1989).

In order to test the statistical validity of these observations, each coefficient was used as a fixed factor in a binomial generalized linear mixed effects model (GLMM) as shown in (7). All four coefficients were combined in the model, and random intercepts and slopes for each coefficient were taken for the random factor of speaker.

(7) Full Model: y = Formality, β = $Coeff_1$ + $Coeff_2$ + $Coeff_3$ + $Coeff_4$,

u = Intercept + Slope of C_1, C_2, C_3, C_4 | Speaker

Null Model: y = Formality, β = C_X + C_Y + C_Z **(Each null model loses a different coeff.)**

u = Intercept + Slope of included coefficients | Speaker

Coefficients: Coeff1 = 0.01 ± 0.06, z-value = 0.12

Coeff2 = 0.53 ± 0.16, z-value = 3.18

Coeff3 = -0.26 ± 0.08, z-value = -3.07

Coeff4 = -0.22 ± 0.09, z-value = -1.98

Random Effects:

Groups	Name	Variance	Std.Dev.	Corr			
speaker	(Intercept)	0.03433	0.1853				
	c1	0.02163	0.1471	-0.32			
	c2	0.16684	0.4085	0.71	-0.52		
	c3	0.01040	0.1020	-0.61	-0.10	0.12	
	c4	0.05501	0.2345	0.79	-0.77	0.91	-0.11

Model Comparison Results: Coeff1: $DF(X) = 1, X^2 = 0.01, Pr(>X^2) = 0.90$

Coeff2: $DF(X) = 1, X^2 = 7.31, Pr(>X^2) = <.01$

Coeff3: $DF(X) = 1, X^2 = 7.42, Pr(>X^2) = <.01$

Coeff4: $DF(X) = 1$, $X^2 = 3.82$, $Pr(>X^2) = <.05$

Stimulus number was not included as a random factor in this model as it did not explain any significant amount of variation in the data, according to model comparison.

The results of model comparison showed that all coefficients were significantly different in informal and formal speech. The visual observation of the lower start point and higher peak of the average informal accentual phrase was shown statistically by the significant difference in coefficients 3 and 4 between speech registers (both significantly higher in informal speech). The significance of coefficient 2 meant that f_0 rose to and fell from the peak more sharply in informal speech which was indicative of greater pitch dynamism in the accentual phrase, an attested indicator of greater overall pitch range (Henton, 1989; 1995). As the functions were normalized for time, no conclusions can be drawn regarding whether the differences in breadth of curvature is related to the actual lengths of accentual phrases in Japanese.

3.7 Discussion

There are a number of points of significance, both methodological and theoretical, which arise from the results of this study. The primary methodological point is that it appears that there are multiple categories of spontaneous speech that researchers much be concerned with when examining variables such as f_0 and articulation rate. Although treating spontaneous speech as a

distinct category is not incorrect as it is phonetically different from read speech (Nakamura et al, 2007), it does appear that simply considering it to be a set category without considering the context of the speech is likely to lead to potential Type I or II errors (depending on the subset of speech being examined) when analyzing the acoustic variables touched on in this study. It is necessary to sub-categorize spontaneous speech both by level of formality, and by whether it is in a conversational register.

There are also some more general points in the realm of phonetics that this study raises. The primary motivation for including f_0 range as a variable in this study, and for the hypothesis that it would be higher in informal speech, was that such a relationship was seen in a previous study on formality in Korean (Winter & Grawunder, 2012). This study found similar results to Winter & Grawunder (2012) where the variable of f_0 range is concerned, which could have a few possible implications; it is possible that rather than only being a property of informal speech in Japanese, an increase in f_0 range is a more general property of such speech cross-linguistically, although it is also possible that it is simply a property of both languages without further implications. Winter & Grawunder also found similar results when testing the relationship between mean f_0 and formality in Korean, again possibly indicating either a broad phonetic similarity between the languages, or a broader cross-linguistic pattern. Further investigation of

this relationship in other languages would be necessary to determine if this is a genuine cross-linguistic property of informal conversational speech.

The results of this study also offer a contrast to previous sociophonetic research on politeness in Japanese. Previous studies have consistently indicated that mean f_0 is higher in polite speech in Japanese (Loveday, 1981; Ohara, 2001), but this runs clearly counter to the results of this study. This contrast is likely due to the differences between spontaneous and elicited speech, and potentially further due to the properties of a more conversational register of speech in Japanese. The results of this study, which in some ways go against the commonly held beliefs regarding the relationship between politeness and f_0 in Japanese, highlight the importance of working with a corpus of natural speech where possible in order to avoid potential type II errors.

Finally, the highly significant relationship between formality and the examined variables in mixed effects models indicates that they could potentially be used by listeners in speech perception when attempting to make category judgments between formal or informal speech. To test this prediction, the study in Chapter 4 will test the salience of the prosody variables investigated in this chapter to the perception of formality in Japanese.

Chapter 4

Prosody and the Perception of Formality in Japanese

4.1 Introduction

4.1.1 Chapter overview

The study described in this chapter investigates the relationship between the prosodic variables examined in Chapter 3 – mean f_0, f_0 range, and articulation rate – and the perception of formality in Japanese. While these variables showed a significant relationship with formality in spoken Japanese, the salience of each individual variable to listeners attempting to determine a speaker's level of formality is still not completely clear. The overall purpose of this study is to determine how much, if at all, these prosodic factors are related to how Japanese listeners perceive different levels of formality.

To investigate this question, an experiment using delexicalized speech stimuli was conducted (see e.g. Pagel et al., 1996; Dellwo, 2008; Morley et al., 2012). Delexicalized speech refers to recordings which are stripped of all phones which a listener could use to determine lexical information, and are left only with prosody (Pagel et al., 1996). This experimental approach has previously been used successfully to test listeners' perception of variation in both f_0 (Morley et al., 2012) and articulation rate (Dellwo, 2008), and so was judged to be appropriate for this project. The stimuli themselves were created based on actual recordings of conversational Japanese speech which were collected for the corpus described in Chapter 3. Further specifics of the process of delexicalization and stimulus preparation will be presented in Section 4.2.2.

The remainder of this chapter will first cover the connection between prosody and both speech perception in general, and as it more specifically relates to the perception of formality or speech register. It will also review some previous work on the perception of prosody in both general synthetic speech, and in delexicalized speech. Section 4.2 will then describe the design of the experiment, and how the stimuli were created. Section 4.3 describes the process of data collection and analysis, and finally Section 4.4 will present a more general discussion of the results and their implications.

4.1.2 Relationship of prosody and formality in speech perception

It has been amply shown in earlier studies of speech perception (such as Abercrombie, 1967; Darwin, 1975; Collier & 't Hart, 1975; Studdert-Kennedy, 1979) that prosody plays a crucial role in perceptual tasks such as speaker identification (Darwin, 1975), pragmatics (Studdert-Kennedy, 1979), and syntactic boundary identification (Collier & 't Hart, 1975). Of greater interest to this study however is the question of whether prosody plays a role in the perception of formality. Previous work has presented a large body of evidence across several languages – including Spanish, Korean, Mandarin, and Japanese – that prosody serves an important purpose in the expression of politeness and formality (see Hidalgo & Cabedo 2014 for an overview), but somewhat less work has been done on precisely how prosody serves to inform our perception and judgment of these phenomena. This section will review two studies which demonstrate that

prosody can be used by listeners to distinguish between different registers of speech both in general, and in Japanese in particular.

The first such study is Laan (1997), which investigated the relationship between prosodic factors including pitch contour, segmental durations, and spectral features on listeners' perception of spontaneous and read speaking styles. The study examined utterances first produced spontaneously by native speakers of Dutch, and then contrasted this spontaneous speech with recordings of the same speakers reading out literal transcripts of their own recordings. Lexically identical sections of spontaneous and read recordings that contained no disfluencies were used as stimuli. The acoustic properties of the stimuli were then artificially manipulated to fall into 5 conditions – one control condition where nothing was changed, and four test conditions: one where the phoneme durations of the two recordings were swapped, one where the pitch contour was replaced with a constant (monotone) f_0, one where the pitch contour was swapped between pairs, and finally one condition where all prosodic factors tested (phoneme duration, f_0 contour, phoneme amplitude) were swapped between pairs. Prosodic manipulations were accomplished using a custom computer program which made use of a TD-PSOLA algorithm (Moulines and Charpentier, 1990). Although the experimental design is not entirely identical to the one used in the current study (see Section 4.2.1 for details), the fact that Laan (1997) investigates the ability of listeners to categorize different speaking styles based largely on prosodic factors makes it an

excellent point of comparison for the current study, and could help set expectations for the results. The results of a perception experiment where subjects were asked to classify recordings as either spontaneous or read showed a significant change in listener responses for each of the test conditions. More specifically, the switching of phoneme duration and pitch contours between the stimulus pairs resulted in subjects' categorization accuracy decreasing, although no single manipulation caused subjects to consistently place stimuli in the opposite category. However, when all three prosodic variables were manipulated (the final test condition), subjects selected stimuli as members of the opposite category at a rate significantly greater than chance, as expected. In terms of the current study, these results indicate that although we should expect prosodic factors to have a relationship to listeners' categorization of speech styles, it is not clear whether or not the manipulation of a single prosodic variable will be sufficient to alter their perception consistently from one category to another, or whether it will simply make the stimuli highly confusable. It does appear, however, that we should expect a significant change in listener categorizations when a number of prosodic variables are manipulated together.

The second study of note is Ito (2001), which investigated the relationship between prosody (f_0 and speech rate in particular) and listener judgments of formality in Japanese. The approach taken by Ito (2001) was to obtain speech materials via a map task (Anderson et al, 1991) with the two speakers in different social positions (one 'higher', one 'lower') in order to obtain

both formal and informal speech. The actual stimulus objects for the perception experiments were only recordings of the word "wakarimashita" ("I understand"), as well as examples of the word read from a script for use as a control condition. Although the approach taken in Ito (2001) is very different from that of the current study – the speech used as stimuli contained lexical information, and there was no manipulation of prosody – the general aim of the study is still similar enough that it can provide us with a useful baseline of expectations for the results of the current study. In Ito (2001)'s perception experiment, subjects were asked to categorize the recordings they heard as either more or less formal than the control condition using the Magnitude Estimation method (Bard et al, 1996), where a response below 1 indicated a more informal stimulus, and vice-versa. Although the results were not entirely consistent, they showed a negative correlation between f_0 and formality judgements, suggesting that a higher f_0 led listeners to classify utterances as more informal. For speech rate, however, there was only a negative correlation with formality for listeners' judgments of one of the two speaker's utterances – with the other speakers' utterances showing a small positive correlation – indicating that speech rate might not be as reliable an indicator of formality as f_0. From this, we might expect to see a stronger relationship between the manipulation of f_0 and listener judgements of formality in the current study than we would with articulation rate.

4.1.3 Perception of synthetic speech

A final topic of note for this chapter is previous work on the perception of synthetic speech, both more generally, and on delexicalized speech in particular. While research has been conducted on the perception of synthetic speech since shortly after the invention of the spectrogram (see e.g. Cooper et al, 1952; Flanagan & Saslow, 1958), of greater relevance to the current study is previous work that addresses the perception of speech created with formant synthesizers (Klatt, 1980) as that is what was used to create the stimuli for the experimental portion of this study (see Section 4.2.2 for further details). Also of interest are studies which investigate the perception of either f_0 or articulation rate in delexicalized speech. This section will outline a number of such studies, as well as their implications for the experimental design discussed in Section 4.2.1.

4.1.3.1 Synthetic Speech

While a significant amount of work has been done on the perception of synthetic speech produced by rule-based synthesizers (see Pisoni, 1997 for an overview), for this study we are primarily concerned with how changes in supra-segmentals (e.g. f_0, articulation rate) may affect listeners' perception of synthetic speech, or by the same token how the fact that the speech is synthetic may affect the perception of the prosody of a recording. An early study which investigated these issues was Klatt (1973) which sought to examine whether the Just Noticeable

Differences in f_0 – defined as the smallest change in f_0 that can be accurately perceived by a listener – is significantly different in instances where the pitch contour is artificially manipulated in synthetic speech as compared to un-manipulated f_0. In this experiment, synthetic examples of the vowel /ɛ/ and the syllable /ja/ were created with the f_0 parameter manipulated in pairs of each stimuli to either be monotone (not moving up or down), both decreasing with an equal slope, both decreasing but with different slopes, and finally with one increasing and the other decreasing. Flanagan & Saslow (1958) previously found a JND in f_0 of ~0.3-0.5 Hz in synthetic vowels, and Klatt (1973) replicated this finding in the condition where f_0 did not increase or decrease. However, the JNDs were significantly higher in the other test conditions, reaching 4.0 Hz in cases where the slope of the change in f_0 was high (32 Hz). This result presents a possible issue for the design of the stimuli used in the current study – although for the most part the natural pitch contours are maintained in the synthetic stimuli in the current experiment, there are some instances where gaps in voicing are connected by linear f_0 ramps (see Section 4.2.2 for further details). If these linear ramps make it more difficult for listeners to perceive changes in f_0 in synthetic speech, it is important to be sure that any changes in f_0 to be examined as part of the current study are greater than the JND of 4.0 Hz found in Klatt (1972). This was taken into account in the manipulations made to f_0 described in Section 4.2.2.

A second study of interest, which investigated how changes in both speech rate and pitch

contour might affect the perception of synthetic speech, was Slowiaczek & Nusbaum (1985), in which subjects were asked to transcribe recordings of varying lengths and syntactic structures which had been generated by a text-to-speech system. Critically, the recordings were synthesized at both slow (150 words/minute) and fast (250 words/minutes) speech rates, both natural pitch contours, and flat (monotone) pitch, the combination of which resulted in 4 test conditions. Although Slowiaczek & Nusbaum (1985) varies from the current study in that it was seeking to investigate the perception of lexical information, while the current study uses delexicalized speech, the results are still of interest as they could indicate whether manipulation of the prosody of synthetic speech could cause unexpected difficulties for listeners. The results of Slowiaczek & Nusbaum (1985) indicated that in syntactically complex sentences, an increased speech rate would have a significant negative effect on transcription accuracy ($p < .001$), while maintaining a natural pitch contour would significantly ($p < .01$) offset this decrease in accuracy in recordings with a fast speech rate. Although listeners in the current study will not be asked to perceive lexical information, the fact that an increase in speech rate appears to cause an increase in perceptual confusability indicates that we might expect subjects to have a somewhat more difficult time judging formality when articulation rate is increased. It also suggests that it will be important to maintain a natural pitch contour as much as possible when creating the delexicalized synthetic stimuli, in order to hopefully ease the difficulty of the perceptual task.

4.1.3.2 Delexicalized Speech

A final important point of background for the current study is whether listeners can make meaningful judgments of variation in prosodic factors such as f_0 and articulation rate in delexicalized speech, and whether they can use this variation to make further judgments about the recordings. One study which investigates such questions is Morley et al (2012), which used a Linear Alignment Model (van Santen & Mobius, 2000) to synthesize delexicalized recordings containing natural-sounding f_0 contours to be used in a perception experiment investigating whether subjects could use synthetic f_0 in delexicalized speech to identify the speakers of the recordings. The experiment involved first playing the subject examples of a natural recording of a given speaker followed by examples of delexicalized speech with similar or different pitch contours, and then asking them whether the delexicalized recording was produced by the same or a different speaker. The results of the study showed that listeners were able to use synthetic f_0 in delexicalized speech to identify speakers at a rate significantly better than chance ($p < .01$) for speakers of both genders, although they were more accurate in the identification of female speakers than males. This result is a positive indication for the current study, as it suggests that listeners can indeed both perceive variation in overall f_0 in delexicalized speech, and they can use it almost as effectively as they could f_0 in natural recordings to make categorization judgments.

A final study of interest is Dellwo (2008), which investigated the salience of speech rate

to listeners' ability to distinguish between different rhythmic classes (i.e. stress- vs. syllable- timed languages), as described by Grabe & Low (2002). The study contained two experiments, of which the second is relevant here. The perception experiment in Dellwo (2008) investigated whether differences in speech rate in delexicalized speech have a relationship with how listeners classify the recordings into either speech with rhythm that was 'regular' (exemplifying syllable-timed languages) or 'irregular' (exemplifying stress-timed languages). Stimuli for the experiment were selected from a group of recordings of read speech in German and French, with the following criteria: the stimuli consisted only of intonational phrases surrounded by pauses, and were selected only if they were one of the 4 qualifying IPs with both the highest and lowest articulation rates from the groups of recordings from each of the two languages, in addition to 4 other randomly selected IPs from each language. Results of the experiment showed that the articulation rate of CV clusters was a strong predictor of listener response, with higher rate correlating with recordings being rated as more regular ($R^2 = .655$, $p < .001$). In terms of the current study, this result indicates both that variation in the articulation rate of CV groups can be perceived by listeners in delexicalized speech, and that it can correlate with differences in listener judgments.

4.1.4. Research Questions and Hypotheses

As discussed in Section 4.1.1, the primary research question addressed by the current study is how much, if at all, the prosodic variables of mean f_0, f_0 range, and articulation rate are related

to the perception of formality in Japanese? More specifically, with the aim of refining the analysis from Chapter 3, a further question examined by this study is how the results of the study might help inform the structure of a predictive statistical model of the relationship between prosody and formality in Japanese? These questions are of interest both as a means of broadly evaluating the salience of prosody to the overall linguistic realization of formality in Japanese, and as a basis for the development of aforementioned statistical model.

The study was designed to test two hypotheses. The first hypothesis is that there will be a correlation between subjects' judgments of the formality of delexicalized recordings and the previous judgments (described in Chapter 3) of the formality of the original recording used to generate the synthesized audio file. Although this hypothesis is not the primary focus of this study, it would nonetheless be notable if listeners are indeed able to judge (non-manipulated) delexicalized speech as formal or informal based solely on prosody in line with how they were judged when lexical information is available. This would imply that, in a general sense, prosody can be used by Japanese listeners in speech perception to help them assess the formality of a given utterance. The second hypothesis this study will test is that the manipulation of the prosodic variables mean f_0, f_0 range, and articulation rate will have a predictable relationship with a change in listeners' judgments of formality. More specifically, this study tests the prediction that manipulating these variables up or down by 20% will cause listeners to judge a recording as more

informal or formal respectively. This is based primarily on the results described in Chapter 3, where those prosodic variables were found to be significantly higher in informal utterances.

To summarize, the hypotheses tested by this study are as follows:

(1). Listeners will be able to judge the formality of an utterance using only prosodic information.

(2). Manipulation of prosodic variables will result in a predictable change in listener judgments of formality.

4.2 Experiment and stimulus design

4.2.1 Experimental design and presentation

In order to test the hypotheses described in Section 4.1.4, a speech perception experiment was designed using Octave (Eaton et al., 2015) which attempts to isolate and test the relationship of each prosodic variable with the subjects' perception of formality, as well as when the variables are changed in combination. To accomplish this, a program written in GNU Octave presented subjects with a randomly ordered set of synthetic, delexicalized auditory stimuli (specifics of the design of these stimuli are covered in Section 4.2.2) which they were asked to judge to be either formal or informal. The script collected their responses via a forced-choice task given on a 6-point Likert-type scale (see Schütze & Sprouse 2014 for an overview of these data collection

methods). The task was presented in this format so that differences in subjects' responses based on the manipulation of the prosody of a stimulus object as compared to a base non-manipulated counterpart could be easily measured and analyzed, in order to better test hypothesis (2). There was also an even distribution of stimulus objects which were based on recordings judged to be formal or informal, in order to help test hypothesis (1) (see Section 4.2.2 for further details). The scale itself was presented on screen in the experiment as shown in (1), and input was via the numeric keypad.

(1).

Informal	Probably Informal	Maybe Informal	Maybe Formal	Probably Formal	Formal
1	**2**	**3**	**4**	**5**	**6**

This scale indirectly asks the subjects to rate their level of confidence in their response (maybe/probably/unqualified), which is done as an attempt to infer the magnitude of the effect of the manipulation of prosodic variables on subjects' perceptions rather than simply forcing a binary choice between formal/informal.

The subjects were also consistently presented with the following information: firstly, before testing began subjects were informed that the auditory stimuli used in the experiment were created based on recordings of speakers of Tokyo-area Japanese. This was genuinely the case, but subjects were made aware of it specifically because listener expectations of speakers have

previously been found to influence their speech perception (see Niedzielski, 1999), particularly as it relates to social information. As f_0 is both a variable of interest in this study, and a common marker of regional dialect in Japanese (Kubozono, 2012) this information was judged to be critical to the subjects' ability to accurately perceive and assess the stimuli. Secondly, before the actual experiment was conducted, and also in order to help establish listener expectations about the auditory stimuli, subjects listened to a single recording of the actual speech of each of the ten speakers whose speech the stimuli were created from. Finally, during the presentation of the stimuli themselves, subjects were presented visually with the age and gender of the original speaker of the delexicalized auditory stimulus. This was done in order to give the subject a general baseline of expectation from which to judge any changes in f_0 and articulation rate.

Subjects were asked to judge a total of 300 stimulus objects, of which 50 were 'base' stimuli, which did not have their prosody manipulated during the synthesis process. There were two different groups of stimuli presented to subjects alternatingly both in order to avoid subject fatigue, and to minimize any potential effects of the ordering of the stimuli, meaning that the experiment made use of a total of 100 base stimulus objects. Each of these base stimuli then had 5 counterparts which had had their prosody manipulated in various ways, resulting in 600 stimulus objects in total. The randomization of the presentation order of the stimuli was handled by a bash script, which assured that no base stimulus object would be encountered either immediately before

or after one of its manipulated counterparts, and that no manipulated stimuli from the same base stimulus object would be encountered in succession. The following section describes the specifics of the prosodic manipulations, and of the process of synthesizing the stimuli.

4.2.2 Stimulus design and creation

As discussed in Section 4.1.1, the auditory stimuli for this study were a series of delexicalized audio recordings containing only a bare minimum of phonetic information for the subjects to perceive. These stimuli were created based on prosodic information taken from recordings made for the corpus of conversational Japanese described in Chapter 3. The first step for the creation of the stimuli was determining precisely what prosodic information needed to be maintained in order to present perceptually meaningful information to listeners, and extracting that information from the original audio files. In the end, only f_0 and power (amplitude of voicing in dB in this case) were judged as necessary for the stimuli in this study, as f_0 was the only parameter of interest to the experiment, and monotonous amplitude values resulted in very unnatural sounding speech.

The next step in the creation of the stimuli was to make a semi-random selection of 100 recordings from the Japanese conversational speech corpus described in Chapter 3 from which to create the synthetic stimuli. This selection was only semi-random as the included recordings, and the selection itself had to meet a number of criteria demanded by the experimental design. Firstly,

the selection had to be made up of an even distribution of recordings from male and female speakers in order to avoid any possible confounds based on speaker gender. Secondly, as mentioned in Section 4.2.1, in order to simplify the testing of hypothesis (1) and avoid any confounds related to having a greater number of stimuli based on one level of formality or another, the selection also had to contain an equal number of recordings which had been previously judged as formal and those judged as informal (see Chapter 3 for more details on how the recordings were assessed). This lead to a set of recordings made up of: 25 male, informal; 25 female, informal; 25 male, formal; 25 female, formal. The recordings were only included in the selection if they were at least 2.5 seconds long in order to both ease the difficulty of the subjects' task somewhat by providing them with at least 10 moras (based on the mean articulation rate seen in Chapter 3) worth of prosodic information, and to maintain consistency in the stimuli themselves. An upper bound for stimulus length was not set, but the possible effect of overall stimulus duration was taken into account in the statistical analysis of the results (see Section 4.3 for more details).

Once the pool of stimuli was created, phonetic information was measured for each recording using the *get_f0* and *pwr* functions of the Entropic Signal Processing System (Entropic Speech Inc., 1989) – or ESPS – in conjunction with bash scripts, at intervals of 10ms. The *get_f0* function was parameterized based on the f_0 information (mean, standard deviation) for each speaker, which was previously calculated for the study in Chapter 3, in order to increase the

accuracy of the f_0 tracking. The f_0 measurements also underwent the following post-hoc manipulation: firstly, due to some jitteriness in the dB values given by *pwr*, the values were smoothed using a running average. This entailed replacing each value with the mean of the value itself and the two following values, and resulted in more natural sounding recordings after synthesis. Additionally, in order to conform to the expectations of the synthesizer used in creating the stimuli, *pwr* measurements were scaled linearly to a 0-60 dB range with 60 as the maximum dB of a recording. Finally, the f_0 vectors were also manipulated; in initial testing of the synthesizing of the stimuli it was found that extended sequences of zero f_0 values would result in extremely unnatural sounding speech. To lessen this effect, vectors were altered using bash scripts which inserted linearly incrementing or decrementing values between the f_0 values at the beginning and end of a series of zero f_0 values. Although this linear change in pitch does not sound entirely natural, it is an improvement over the sudden gaps caused by allowing the zero values to remain. Table 4.1 shows an overview of the phonetic parameters of interest (f_0, articulation rate, f_0 range) of the recordings used to create the synthetic speech.

Table 4.1: *Descriptives of phonetic parameters of the randomly chosen stimuli.*

	Informal		Formal	
Variable	**Mean**	**SD**	**Mean**	**SD**
Mean f_0	167.6Hz	44.6 Hz	150.1 Hz	41.3 Hz
Articulation Rate	7.84 m/s	1.3 m/s	6.22 m/s	1.0 m/s
f_0 Range	135.1 Hz	50.8 Hz	90.4 Hz	39.1 Hz

Table 4.1 shows that, on the whole, the phonetic parameters of the random pool of stimuli pattern quite closely with those seen in Chapter 3. It is therefore reasonable to assume that the results will not be unduly affected by an initial group of stimuli that differ unexpectedly from previously observed patterns.

Once these acoustic parameters were acquired, parameter files were automatically created for use in a command-line Klatt synthesizer (Klatt, 1980; Klatt & Klatt, 1990; Iles & Ing-Simmons, 1994). The Klatt synthesizer allows for the creation of synthetic recordings via the specification of different control parameters corresponding to the acoustic properties of the recording. For example, it is possible to manipulate the f_0, amplitude of voicing, formants, and voice quality of frames of a pre-defined length (15ms-long blocks by default, 10ms frames for this study). Such a synthesizer is ideal for the creation of de-lexicalized speech, as it allows removal of the variation in parameters which would otherwise provide phonological information to the listener (such as e.g. changes in formants for vowels) while maintaining variation in the

parameters of interest (in this case, f_0). In the case of the stimulus objects for this study, F_1, F_2, and F_3 were set to constant values of 500, 1500, and 2500 respectively in order to create the impression of the entire recording being an extended /ə/ (see Klatt, 1980: 986) while maintaining the f_0 and amplitude variation measured previously with ESPS. All other parameters were kept constant, with turbulence and tilt parameterized to produce a slightly creaky voice quality. A full list of the constant values used in the Klatt parameter files can be found in Appendix IV.

Once these parameter files were created for the synthesizer, each one was used to create a series of further parameter files which artificially manipulated the prosodic variables of interest to the study. Each base file was used to generate parameter files for 5 further stimuli, including files modifying each prosodic variable of interest individually in the direction opposite from the recording's original level of formality (i.e. a recording that was originally informal would have mean f_0, f_0 range, and articulation rate each adjusted downwards, which would lead us to expect subjects to judge the resulting stimuli as more formal), and files modifying **all** variables both up and down. The simplest manipulation was to mean f_0, which simply entailed adding or subtracting 20% of the mean of the entire vector from each value, leading to a complete shift of the f_0 contour upwards or downwards. In order to manipulate f_0 range, values were first z-transformed (by subtracting the mean from each value and then dividing by the standard deviation) so that the values were centered on a mean of 0. The values were then each multiplied by 1.2 to increase

range, or 0.8 to decrease range, meaning that values above the mean would be increased, or decreased respectively, with the degree of change being proportional to how far the value was from the mean (i.e. a value farther from the mean would be increased or decreased more). This resulted in the f_0 values expanding away from or towards the mean, resulting in an overall change in the f_0 range. Finally, manipulation of articulation rate was achieved during the synthesis process itself; the audio files, which were originally encoded at 16 kHz, were synthesized at either 12 kHz (for decreased articulation rate) or 20 kHz (for an increased rate), and then re-encoded from raw at 16 kHz using the Unix command line tool SoX in order to force the file to be replayed faster or slower than it was originally. This route was taken, rather than simply speeding or slowing the playback manually, so that the f_0 values in the parameter files could also be manipulated up or down by the same percentage as the sample rate was changed, in order to offset the perceived change in pitch that results from an audio file being played at a changed sample rate.

Finally, the stimuli were automatically synthesized using bash command line tools. The manipulations for articulation rate were applied to parameter files before mean f_0 and f_0 range were altered to allow those manipulations to f_0 to occur after it had already been changed to compensate for the changes in sample rate. The manipulation of the prosodic variables resulted in stimulus objects that were noticeably different from the 'base' stimuli, and in the cases where

all three variables were manipulated, were very difficult to identify as being based on the same source recording based on post-experiment interviews with the subjects.

4.3 Data collection methodology and analysis

4.3.1 Overview of experimental subjects

The experiment was conducted in the Phonetics Laboratory at Oxford University. In total, 16 linguistically naive native speakers of Japanese – 5 male and 11 female – participated in this experiment. There was no requirement that subjects be from a specific region of Japan, as it was assumed that subjects would have a high degree of familiarity with the 'standard' Tokyo Japanese regardless, but the subject pool was limited to people who were born and raised in Japan until at least age 18. Subjects were recruited at both Oxford University and Oxford Brookes University, and included short-term exchange students as well as graduate students and post-docs.

4.3.2 Experimental procedure

Subjects were welcomed to the lab, and then given a brief overview of the experiment which instructed them to expect recordings to be from speakers of Tokyo-area Japanese, and to expect recordings with the actual words obscured. The concept of formality was defined for the subjects for the purposes of this experiment as speech "which appeared to be among friends or colleagues" for informal speech, and "speech towards elders or superiors" for formal speech.

Subjects were then played a sample of a non-synthesized recording from each speaker in the experiment. Finally, subjects were introduced to the experiment itself via a brief practice section (two stimuli were presented). Subjects were given no specific instruction on the use of the six-point scale (as described in Section 4.2.1) other than that their task was to judge each recording as being formal or informal speech. Finally, subjects were informed that the experiment would involve 300 stimulus objects, and that they were allowed to take a break whenever they wished (the experiment was self-paced, and automatically paused after each selection). All instructions were given to the subjects in English, for consistency.

After the experiment was complete, a brief exit interview was conducted with each subject. Subjects were told the general purpose of the experiment, and shown examples of the manipulation of the prosody of the recordings. No subject reported becoming aware of the purpose of the experiment or of the manipulation before it was explained. Some subjects reported that they tended to judge stimuli as more informal when they noticed large rises or falls in f_0 (seemingly representing variation in f_0 range).

4.3.3 Data overview

With each subject judging half of the total pool of 600 stimuli, this resulted in a total of 4,800 responses. There are a few points of note that are apparent from a general examination of the data. First is that the overall mean of the responses falls at 3.38 on the 6-point scale, meaning

that on average there was very close to a 50/50 split in how stimuli were judged (although it does skew very slightly towards subjects judging stimuli as informal). This could be due to the fact that there actually was a 50/50 split of stimuli that were originally of each different formality, or it is possible that subjects were simply making an effort (whether conscious or not) to balance their responses.

In order to determine if there actually is a relationship between the original formality of the stimulus items and subjects' responses, and to thereby also test hypothesis (1), the relationship between formality and subject response was analyzed using the cumulative link mixed model in (2).

(2). Full Model: y = Response, β = Base Formality,

u = Intercept + Slope of Base Formality | Subject

Null Model: y = Response, β = Intercept

u = Intercept + Slope of Base Formality | Subject

Coefficients: BaseFormalityF = 0.2053 ± 0.1287, z-value = 1.595, Pr(>|z|) = 0.11

Threshold Coefficients: 1|2 = -2.92 ± 0.18, 2|3 = -1.10 ± 0.12

3|4 = 0.23 ± 0.11, 4|5 = 1.91 ± 0.13

5|6 = 3.60 ± 0.27

Random Effects:

Groups	Name	Variance	Std.Dev.	Corr
subject	(Intercept)	0.062177	0.24935	
	FormalityF	0.005301	0.07281	-1.000

Model Comparison Results: Base Formality: $DF(X) = 1, X^2 = 2.41, Pr(>X^2) = 0.12$

Here only *subject* is included as a random factor, as this model only tested the relationship between subject responses and a stimulus object's base formality for those objects which had not been manipulated (and thereby retained the prosodic structure of the original recordings). Overall the model did not show a statistically significant relationship, although it **did** show the expected pattern – mean response to unmanipulated stimuli was lower on average, i.e. judged as more informal when the original stimuli was informal, and higher when the original stimuli was formal – but subjects were unable to make this distinction at a rate significantly better than chance in the unmanipulated stimuli.

As the experiment only recorded the ordinal response to each stimulus from the subjects, in order to compare the responses to manipulated stimuli to their respective 'base' stimuli, a script was created to calculate the differences in responses between each base variable and all of its related stimuli. This allows us to test hypothesis (2). Figure 4.1 shows the overall distributions of all of these changes split by the direction of manipulation of the prosodic variables.

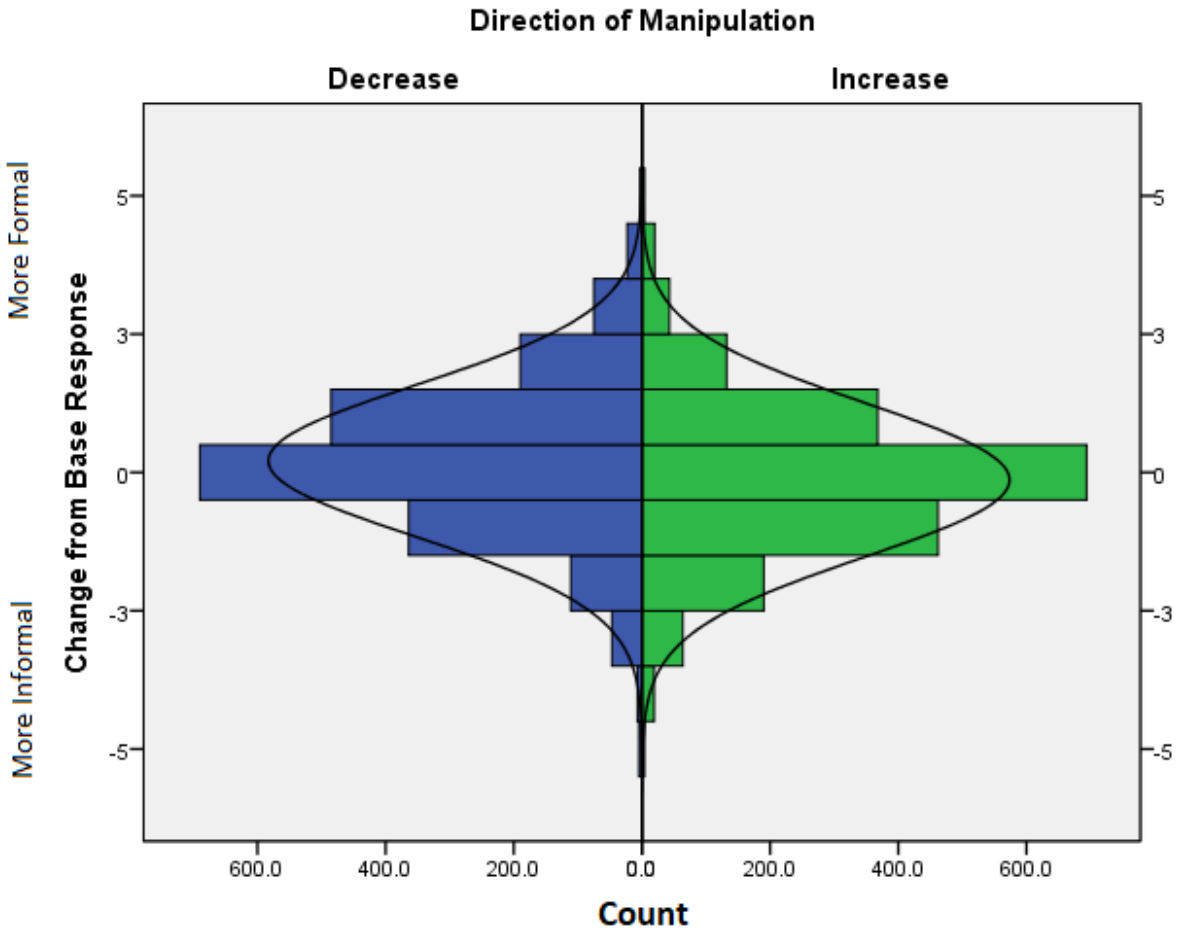


Figure 4.1: *Histogram of the changes in response from the base stimuli to the manipulated stimuli, split by the direction of manipulation of the prosodic variables (whether they were increased or decreased).*

An examination of the histogram shows a pattern emerging. Although the results are not completely consistent and show changes in both directions for both directions of manipulation, the overall trend appears to go in opposite directions based on the split. While there are many tokens with no change for both directions, following what would be expected if hypothesis (2) were true we see a greater volume of tokens with a positive change when the variables were decreased (making them appear more formal), and the opposite is seen when the variables were

increased. This split appears to be slightly more pronounced when the variables were decreased.

In order to further examine this pattern, Figure 4.2 shows a bar graph summarizing how the direction of manipulation for each variable is related to the change in subjects' responses.

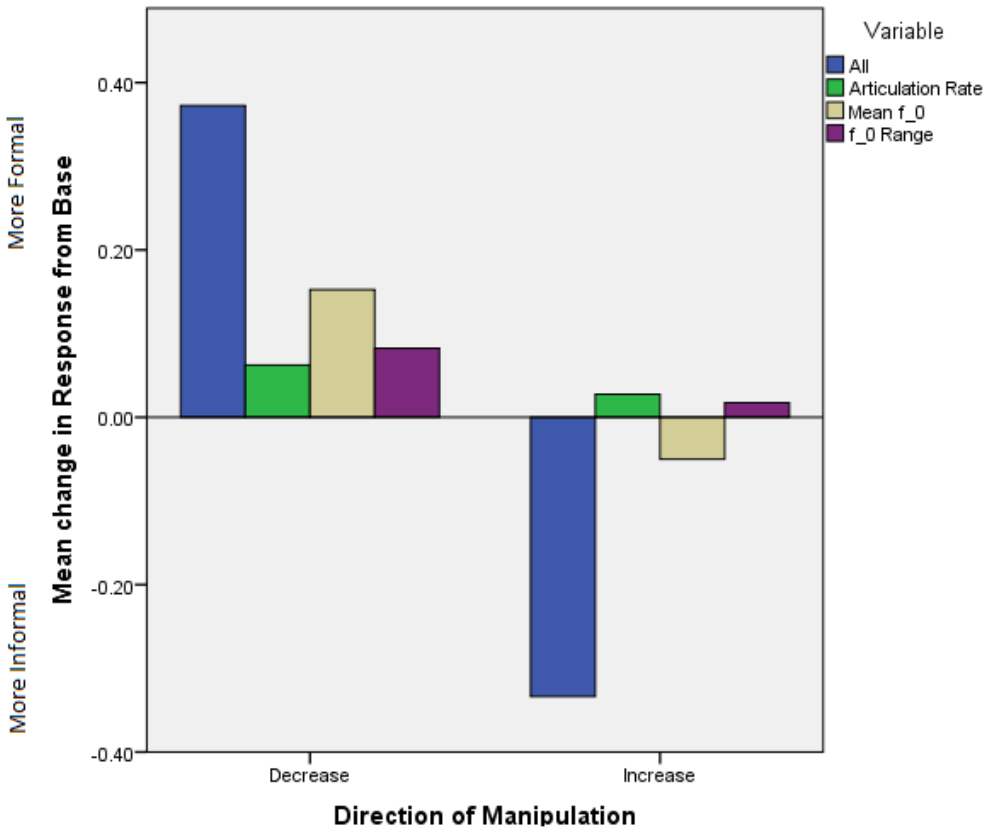


Figure 4.2*: Relationship between the direction of manipulation of each prosodic variable and the change in subjects' responses.*

Based on hypothesis (2), we would expect the change to be positive (judged as more formal) for stimuli where the prosodic variables were decreased, and conversely for the change to be negative when stimuli were manipulated upwards. Visual examination of Figure 4.2 shows that this expected trend appears to hold strongly when all three variables are manipulated together.

However, when the variables are examined individually, the trend only appears to hold strongly in the case of decreases in the variables (which we hypothesize to indicate more formal speech). Only mean f_0 maintains the trend in both conditions, with articulation rate and f_0 range actually going opposite to the expected trend when the variables are increased, although the actual change in these cases is not very large (less than 5% of a step on the Likert scale), and the correlation is not statistically significant. It does appear that in spite of this, however, the effects are roughly equal in opposite directions when all of the variables were manipulated, indicating a deeper story, which will be discussed further in Section 4.4.

4.3.4 Modeling analysis

In order to test the significance of the pattern seen in Figures 4.1 and 4.2, an analysis was conducted using cumulative link mixed models in R (R Core Team, 2017) found in the ordinal library (Christensen, 2015). All of the manipulation conditions were tested individually and in combination using the models given in (3).

(2) a. *All data model*

Full Model: $\boldsymbol{y}$ = Change in Response, $\boldsymbol{\beta}$ = Direction x Base Formality,

$\boldsymbol{u}$ = Intercept + Slope of Direction | Subject, Stimulus & Subtype

Null Model: $\boldsymbol{y}$ = Change in Response, $\boldsymbol{\beta}$ = BaseFormality

$\boldsymbol{u}$ = Intercept + Slope of Direction | Subject, Stimulus & Subtype

Coefficients: DirectionD = 0.797 ± 0.19, z-value = 4.09, Pr(>|z|) < .001

BaseFormalityF = 0.404 ± 0.17, z-value = 2.35, Pr(>|z|) < .05

Direction x BaseFormality = -0.25 ± 0.17, z-value = -1.40, Pr(>|z|)=0.15

Threshold Coefficients: -5|-4 = -5.75 ± 0.37, -4|-3 = -4.38 ± 0.23

-3|-2 = -2.91 ± 0.18, -2|-1 = 1.67 ± 0.17

-1|0 = -0.26 ± 0.16, 0|1 = 1.32 ± 0.16

1|2 = 2.70 ± 0.17, 2|3 = 3.90 ± 0.18

3|4 = 5.16 ± 0.21, 4|5 = 7.15 ± 0.41

Random Effects:

Groups	Name	Variance	Std.Dev.	Corr
stimulus	(Intercept)	3.096e-01	0.5564	
	direction	1.209e-06	0.0011	1.000
subject	(Intercept)	5.973e-02	0.2444	
	direction	1.412e-01	0.3757	-0.791
subtype	(Intercept)	1.906e-02	0.1381	
	direction	8.097e-02	0.2846	-1.000

Model Comparison Results: Direction: $DF(X) = 2, X^2 = 9.46, Pr(>X^2) < .01$

Base Formality: $DF(X) = 2, X^2 = 5.557, Pr(>X^2) = 0.06$

b. *Subtype models*

Full Model: y = Change in Response, β = Direction x Base Formality,

u = Intercept + Slope of Direction | Subject & Stimulus

Null Model: y = Change in Response, β = BaseFormality

u = Intercept + Slope of Direction | Subject & Stimulus

Coefficients:

(All) DirectionD = 1.196 ± 0.21, z-value = 5.69, Pr(>|z|) < .001

(f_0) DirectionD = 0.259 ± 0.17, z-value = 1.51, Pr(>|z|) = 0.13

(Rate) DirectionD = 0.101 ± 0.21, z-value = 0.47, Pr(>|z|) = .63

(Range) DirectionD = 0.072 ± 0.17, z-value = 0.42, Pr(>|z|) = 0.67

BaseFormalityF = 0.353 ± 0.15, z-value = 2.26, Pr(>|z|) < .05

Direction x BaseFormality = -0.26 ± 0.18, z-value = -1.44, Pr(>|z|)=0.15

Threshold Coefficients: -5|-4 = -5.90 ± 0.59, -4|-3 = -4.34 ± 0.30

-3|-2 = -2.85 ± 0.19, -2|-1 = -1.55 ± 0.16

-1|0 = -0.10 ± 0.15, 0|1 = 1.52 ± 0.15

1|2 = 2.91 ± 0.17, 2|3 = 4.11 ± 0.20

3|4 = 5.11 ± 0.25, 4|5 = 7.04 ± 0.52

Random Effects:

Groups	Name	Variance	Std.Dev.	Corr
stimulus	(Intercept)	0.221516	0.47065	
	direction	0.002131	0.04616	1.000
subject	(Intercept)	0.164376	0.40543	
	direction	0.427799	0.65406	-0.963

Model Comparison Results:

(All) Direction: $DF(X) = 2$, $X^2 = 19.60$, $Pr(>X^2) < .001$

($f_0^{\wedge}$) Direction: $DF(X) = 1$, $X^2 = 2.20$, $Pr(>X^2) = 0.13$

(Rate) Direction: $DF(X) = 1$, $X^2 = 0.23$, $Pr(>X^2) = 0.63$

(Range) Direction: $DF(X) = 1$, $X^2 = 0.17$, $Pr(>X^2) = 0.67$

Base Formality: $DF(X) = 2$, $X^2 = 4.98$, $Pr(>X^2) = 0.08$

These models bear some further discussion. Subtype here refers to the phonetic parameter that was manipulated (either all, mean f0, articulation rate, or f0 range), and was included as a random factor in only the model of the full data set. An interaction between direction of manipulation and the base formality of the stimulus was included to test if subjects were more likely to change their responses based on the original formality of the recordings. Subjects did appear to be slightly more likely to change their responses to a more formal one when the base stimuli's formality was

formal (and vice versa), but this effect was not significant according to model comparison. The fixed factor coefficients given refer to the point of maximum likelihood on the normalized response scale given by the threshold coefficients (where, for example $1|2$ is the coefficient at which a response of 1 or 2 are equally likely on an ordinal scale). In relation to this experiment, a positive fixed effect coefficient means that the change was more likely to be towards the 'formal' end of the original Likert scale. Table 4.2 shows a simplified overview of the results of the modeling analysis.

Table 4.2: *Overview of modelling results.*

Variable	$X^2(2)$	Estimate	p-value
Full Data Set	9.46	$.797 \pm .19$	$p < .01$*
All	19.60	$1.169 \pm .21$	$p < .001$*
Articulation Rate	.23	$.101 \pm .21$	$p = .63$
Mean f_0	2.20	$.259 \pm .17$	$p = .13$
f_0 Range	.17	$.072 \pm .17$	$p = .67$

The results of the modelling analysis agree with what can be seen in Figure 4.2. The variables tested show no significant relationship with direction of manipulation when tested individually. However, when all three variables were manipulated the relationship is significant, with the models showing that change in response was significantly higher (towards the formal side) when the variables were manipulated downwards, as hypothesized.

4.4. Discussion

Overall, the analyses presented in Sections 4.3.3 and 4.3.4 appear to lend qualified support to both of the hypotheses in Section 4.1.4. Hypothesis (1) is less supported, as although there is a relationship between the formality of the recording the base stimuli were created from and the subjects' judgments of the formality of those stimuli as predicted, this relationship falls short of statistical significance. This is likely indicative of the sheer difficulty of the task of judging the formality of delexicalized recordings, and the fact that a pattern appears to exist at all independent of any manipulation is interesting, indicating that listeners might perform better if given more information (such as being allowed to listen to more examples of the original speakers). Another point worth noting is that it appears that subjects had a harder time judging base stimuli which were modeled on informal recordings, as can be seen in the boxplot in Figure 4.3.

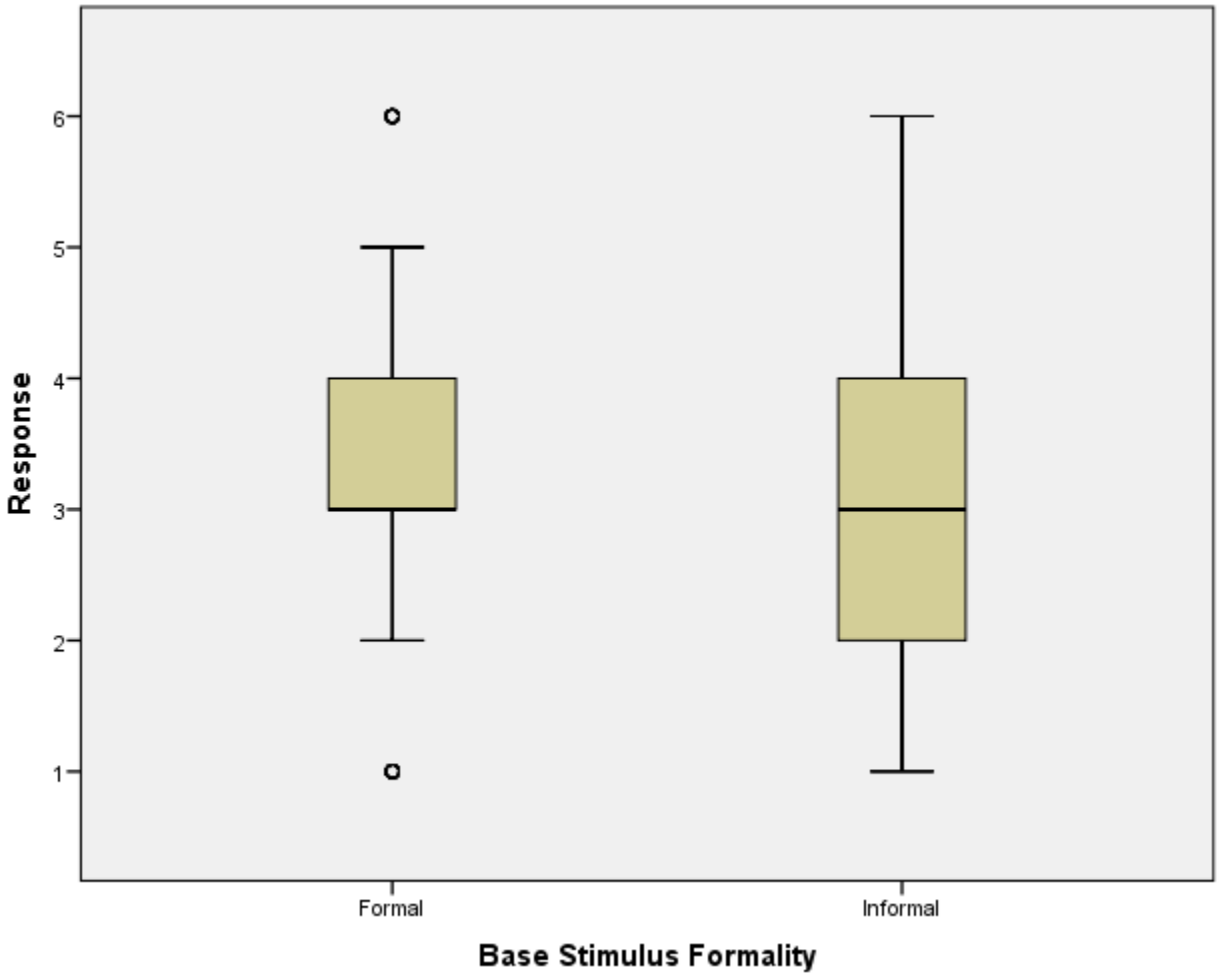


Figure 4.3: *Boxplot of subject responses split by the formality of the original recording.*

Although the overall trend for informal base stimuli does lean slightly towards subjects judging more of them to be informal, there is an almost equal split in cases where they were unsure (responses 3 or 4 on the scale), which is not the case when the stimuli were originally formal. This provides some suggestions as to the possible reasons for the patterns found when examining the changes in responses from these judgments of the base stimuli.

As discussed in Section 4.3.3, there is a significant relationship between the direction of the manipulation of prosodic variables and the changes in subject responses as predicted in

hypothesis (2). However, this pattern only appears to hold consistently when all three variables are manipulated, and particularly breaks down when the individual variables are manipulated in a direction hypothesized to make them appear more informal. This follows the results of Laan (1997), in that the results appear to be *super-additive*, meaning that the total effect of manipulating multiple variables is greater than it would be if you simply added the effects of manipulating each variable individually. This discrepancy, as well as the fact that subjects appeared to have a harder time judging informal base stimuli as such, could be at least partially explained by theories of speech perception espoused in e.g. Norris & McQueen (2008) and Kleinschmidt & Jaeger (2015), where perceptual categories (e.g. vowels, syllables, social categories) are realized as distributions (often Gaussian in shape) of phonetic values. To illustrate this concept as it applies to the current study, Figure 4.4 shows a series of hypothetical normal distributions of f_0 values from an imaginary speaker.

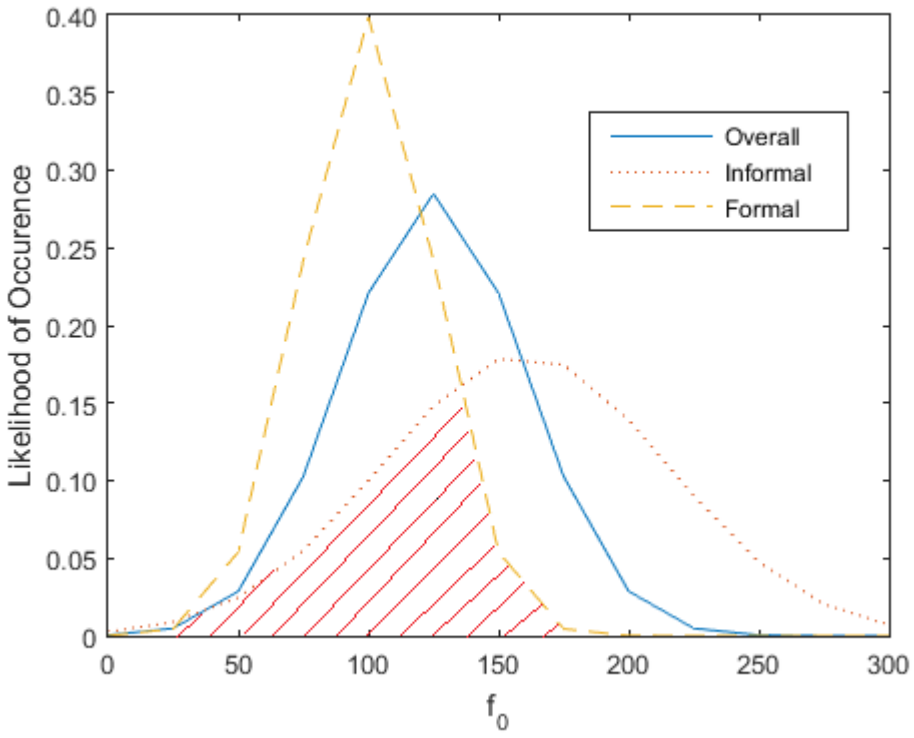


Figure 4.4: *Hypothetical distributions of a speaker's f_0. Shows possible distributions of f_0 for a speaker's overall speech, formal speech, and informal speech. The red highlight shows the overlapping area of the distributions for formal and informal speech.*

Figure 4.4 contains three distributions: first, an overall distribution of f_0 values for the speaker. This encompasses all of the speaker's hypothetical utterances. The other two distributions, however, are meant to represent possible distributions of f_0 values in both formal and informal speech. Based on the results of the study in Chapter 3, the distribution for informal speech has a both a higher mean (160 Hz), and a higher standard deviation (55 Hz, representing f_0 range). This results in a notably wider distribution for informal speech than for formal (mean of 100 Hz, SD of 25 Hz), and also results in an area of overlap between the two distributions.

The general hypothesis in such theories of speech perception is that if a particular

perceptual item falls in an area covered by both distributions, there will be a greater probability of it being selected as a member of the narrower distribution (Kleinschmidt & Jaeger, 2015), the peak of which is more likely. This can be seen by examining the y-axis of Figure 4.4, which represents the likelihood of a token at a given point on a distribution occurring in that category. This means that, for example, if a given utterance had a mean f_0 of 125 Hz, it would have a notably higher likelihood of occurring in the formal distribution. If this is the case, it would then follow that in cases where the distributions overlap, listeners are more likely to judge a token that falls into that part of the overall distribution as a member of the narrow (in this case formal) distribution. Therefore, it is only when the relevant variables are pushed to higher levels that a listener would become more likely to judge a token to be a member of the wider distribution (informal, in the case of this study).

If this is the case, it would provide an explanation both for why informal tokens were more difficult to categorize, and why the hypothesized pattern appears to hold better when variables were manipulated in a direction expected to be perceived as more formal. As in the example in Figure 4.4, decreasing the variables from the overall mean would lead to a point where it is much more likely to occur as part of the narrower formal distribution, while conversely increasing the value initially leads to points where category membership is still somewhat ambiguous. Essentially, there is a larger section of the distribution of informal utterances where

a token is likely to be confused for a formal utterance based on phonetic information alone than there is of the distribution of formal utterances, which could at least partially explain subjects' added difficulty in judging informal stimuli. Assuming that the distribution of phonetic values representing "formality" to a speaker of Japanese is multi-faceted (as it appears to be, based on the results of the studies in this thesis), and that the possible distribution representing informal utterances is wider than that representing formal utterances, it is logical that it would be more necessary to increase *all* the variables rather than only one in order to push the stimulus to a point on the distribution where it is easier for a listener to correctly identify an informal utterance as such. This could be tested in the future by running a similar study to the one described in this chapter, but at the same time having a more pronounced increase (perhaps double the increase as for the stimuli in this study) in the prosodic variables of interest when they are moved in that direction.

One further point to note regarding this hypothesis, however, is that due to the super-additive nature of the results it is also possible that listeners are not using any one variable in isolation, and therefore any explanation focusing on distributions of a single variable may be somewhat reductive. While the results do lend a certain amount of support to a theory where speakers are accessing cue distributions similar to those seen in Figure 4.4, a model of this category judgment task must necessarily be based on a ‘combined’ distribution of all the relevant

phonetic parameters considered together.

This hypothesis that levels of formality in spoken Japanese can be represented as distributions of phonetic values presents a clear way forward towards a probabilistic model. As Bayesian models are based largely on probability distributions (Nicenboim & Vasishth, 2016), and can also incorporate multiple variables together into the model, they appear to be an extremely appropriate tool for modelling the production and perception of formality in Japanese. Chapter 5 will cover the theoretical basis and creation of such a model in a Bayesian framework.

Chapter 5

Modeling Formality in Spoken Japanese Using Bayesian Inference

5.1 Introduction

5.1.1 Chapter overview

Based on the significant relationships between the prosodic variables of mean f_0, articulation rate, and f_0 range found in the experiments in previous chapters, this chapter describes a probabilistic statistical model which uses information on the prosody taken from the corpus of spoken Japanese described in Chapter 3 to predict the level of formality of a recording of spoken Japanese independent of any lexical information. Further information on the specific goals of this model is presented in Section 5.1.2.

The model was created by making use of Bayes' Rule (Bayes & Price, 1763) and Bayesian Inference in a framework similar to an *ideal observer model* (Geisler, 2003). This is not a trivial task, as the experiment in Chapter 4 demonstrated that even human listeners have a difficult time classifying recordings as formal or informal given only prosodic information unless a number of prosodic variables are manipulated to a point where the decision becomes more clear-cut. This means that while it is likely not realistic to expect the model to predict categories with a high level of accuracy, it is still reasonable to expect the prosody of recordings to have some degree of predictive power.

The remainder of this chapter will outline the goals and motivations behind the model, and will then give a brief overview of the two core concepts on which the model is based:

Bayesian statistics and ideal observer models. It will then describe in detail the structure of the model and the rationale behind its design. Results of the predictive model will then be described, along with further analysis of what these results indicate as they relate to previous results and future research. Finally, there will be a discussion of the implications of the model for theories of speech perception, and for the study of meta-linguistic information both in Japanese and more generally.

5.1.2 Modelling Goals

Before continuing on, it will first be useful to give an overview of the specific goals of the model being described in this chapter, in order to better motivate both the decision to make use of Bayesian inference and the design decisions of the model.

In computational terms, the model has two main goals: the first is, as has been stated, to predict the formality of a recording based on prosodic information. The second computational goal of the model is to represent the super-additive nature of the results – this means that any model will be looking at all of the predictive phonetic properties as a whole, rather than considering them individually. These computational goals are fairly straightforward, and could theoretically be accomplished using a few different statistical approaches, including a (frequentist) logistic regression, or a linear discriminant analysis (Fisher, 1936). The results of modelling the data under both of these frameworks will be touched upon briefly in this chapter.

In addition to the computational goals, an additional structural goal of the model is to simulate as closely as possible the way in which human listeners performed the category judgment task (as in Chapter 4). This means that the model must have a few specific structural properties – it must, based on the conclusion of Chapter 4, represent the prosodic cues collectively as probability distributions, and must also provide a means of quantifying the inherent uncertainty of the categorization task. Additionally, it should be able to incorporate information about the prior knowledge and expectations of the human listeners, which was gained from the experiments in Chapters 3 and 4. Taking all of these goals into account, the Bayesian framework appears to be the most structurally appropriate approach to this modelling task, as it has the ability to incorporate prior information, and its general computational structure also patterns closely to the theory of the cognitive process behind the categorization task as described in Chapter 4.

The overall goal of creating this model and structuring it as closely to our idea of the cognitive processes underlying phonetic categorization tasks as possible is to fill in some gaps in the results of Chapter 4, and to provide a baseline of expectations for future research on the topic. Participants in the study in Chapter 4 were given very minimal information on the speakers of the delexicalized speech stimuli, and had to rely largely on their general prior knowledge of Japanese to make judgments of formality. It is possible that this was one of the main factors that lead to the subjects having difficulty accurately categorizing the unmanipulated stimuli, and so if

this model – which will take into account information on individual speakers which the human listeners did not have – is a reasonable representation of the category judgment task then it should be able to provide a better 'real world' baseline of expectation for categorization accuracy in humans when they have actual information about who they are listening to.

5.1.3 Bayesian statistics

At their most basic, Bayesian statistics involve the calculation of *posterior probabilities*, which are the probabilities of a hypothesis given some information about the *prior probability* of the hypothesis, and the *likelihood* of the hypothesis given the observed data. This process is expressed by Bayes' Rule (Bayes & Price, 1763), as shown in (1).

(1) $$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}$$

In plain language, this can be read as: the probability of A given the data B is equal to a proportionality of the Probability of the observed data B given A (the likelihood) times the prior probability of A. The denominator of the equation is generally calculated as a sum of the probabilities of each category. As the equation is a proportionality, it is often written in simpler form as in (2).

(2) $Posterior \propto Likelihood \times Prior$

This rule is often illustrated via the example of LaPlace's Demon (LaPlace, 1820). In this example, we attempt to calculate the probability of going to hell after consorting with the demon

LaPlace, given some prior knowledge about the overall probability of going to hell, and some knowledge of the likelihood of people who have been sent to hell having consorted with the demon. We have the following information: 75% of the population goes to hell, while 25% goes to heaven. In a smaller observed sample, out of 9 people sent to hell, 6 had consorted with the demon ($= 66.6\%$), while out of 7 people sent to heaven 5 had consorted ($= 71.4\%$). So, we know that the prior probability of going to hell is 75% independent of any other factors. Also, given observation of a subset of the people sent to hell, we see that the likelihood of having consorted with the demon among those sent to hell is 66.6%, while among those sent to heaven it is 71.4%. We can then plug these probabilities into Bayes' rule to calculate the posterior probability of going to hell after consorting with the demon, as in (3).

(3) $$Pr(Hell|Consort) = \frac{0.666 \times 0.75}{0.666 \times 0.75 + 0.715 \times 0.25} = 0.737$$

We can see that the posterior probability of going to hell after consorting with the demon is 73.7%, actually lower than the prior probability of 75%.

More recently, the general framework of Bayes' rule has been extended to be usable in more forms of statistical analysis, including in particular in the evaluation of hypotheses given the observation of continuous data rather than simply discrete probabilities. This type of statistical modeling is referred to as Bayesian inference (Box & Tiao, 2011), and involves calculating posterior probabilities based on observation of *distributions* of the parameters of interest in a data

set. While a full overview of the mathematics underlying Bayesian inference is beyond the scope of this chapter, there are a few general points of note.

One of the main differences between frequentist statistics (such as t-tests, ANOVAs, or mixed models) and Bayesian inference lies in the nature of how the results are presented and interpreted. In frequentist statistics, the results are in the form of some discrete statistic (such as F values for ANOVAs, or t values for mixed models), and a P value is an expression of the probability of observing an equal or more extreme statistic were the null hypothesis true (Niceboim & Vasishth, 2016). Bayesian P values, on the other hand, are derived from a posterior distribution of the probability of the hypothesis being true given the data. Therefore, unlike frequentist P values, Bayesian P values relate directly to the probability of a null hypothesis being true. This is useful to the model in this chapter mainly because it allows the model to simulate the uncertainty inherent in the prediction it is attempting to make by observing the posterior probabilities and making use of different points in the distribution rather than simply a single value. This aspect of model design will be discussed further in Section 5.2.

Another important point regarding the structure of Bayesian inference relates to how Bayes' rule can be used to express the task of inference under uncertainty, which is essentially what a listener is doing when they attempt to make a decision regarding the category membership of an observed sound or utterance given some prosodic information (Kleinschmidt & Jaeger,

2015). This concept of inference under uncertainty can be described in the Bayesian framework using ideal observer models.

5.1.4 Ideal Observer Models in speech perception

An ideal observer model is a statistical model which attempts to represent the cognitive task of perceptual discrimination between categories in a Bayesian statistical framework (Geisler, 2003). Under this framework, the *likelihood* in Bayes' rule relates to an observer's mental representation of the likelihood of cues to a given category in a particular situation, while the *priors* relate to an observer's mental representation of the possible cues to a given category in general. An early use of this framework was in automatic speech recognition (see e.g. Baker, 1979; Bahl et al., 1983; Goel & Byrne, 2000) and it has recently been adapted for use in speech perception research as an *ideal listener model* (see e.g. Clayards et al., 2008, Norris & McQueen, 2008; Kleinschmidt & Jaeger, 2015), where the distributions of the likelihood and prior probabilities are particularly related to representations of the likelihood of *acoustic* cues to the category membership of an observed sound or utterance. This ideal listener model for predicting category membership is formalized in Kleinschmidt & Jaeger (2015) as in (4).

(4) $$p(C = c_i|x) \propto p(x|C = c_i)p(C = c_i)$$

Where here $C = c_i$ refers to the probability of membership in a given category, the likelihood refers to the probability of the observed data x given category c_i, and the priors refer to the prior

probability of category c_i in all of a speaker's previous observation of the given acoustic cue x. Of primary importance to the model in this chapter is the likelihood, which in terms of speech perception specifically refers to an observation of an acoustic cue in a particular situation in which a listener is attempting to make a category judgment. In terms of our model this would be the cue distributions of a particular speaker, but it could also refer to other categories such as social group, age group, or gender. This ideal listener model expresses the theory, previously discussed in Chapter 4, that listeners access distributions of acoustic cues when making judgments in speech perception, and also that listeners develop cue distributions for particular speakers or speaker groups. This concept is illustrated in Figure 5.1, from Kleinschmidt & Jaeger (2015).

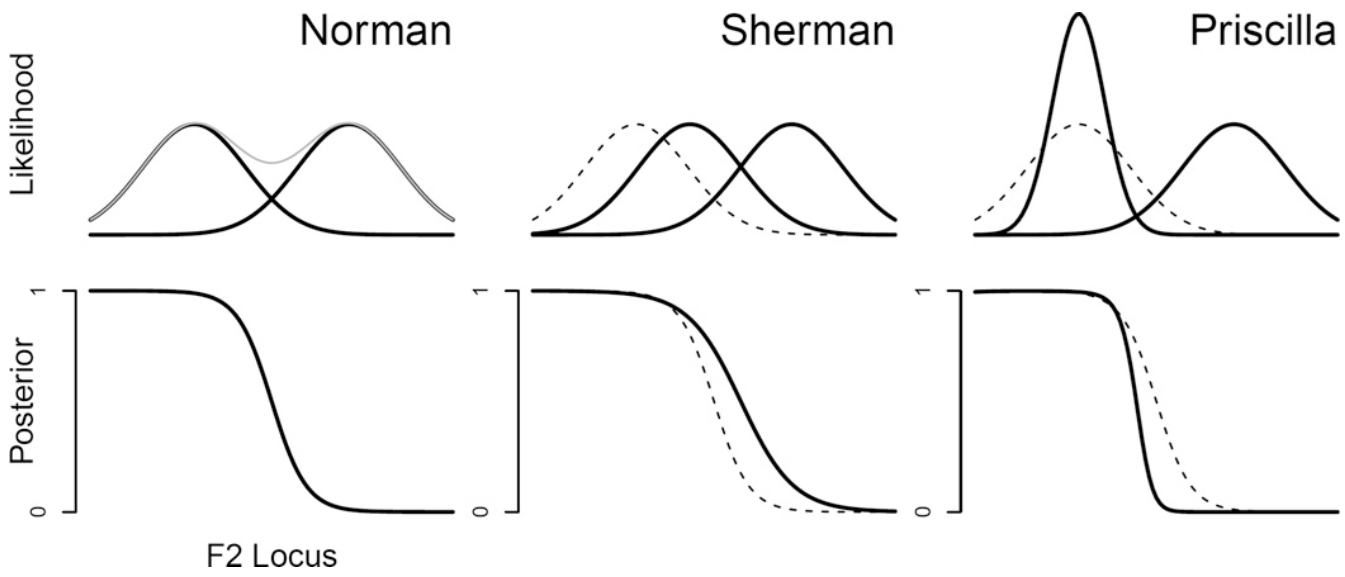


Figure 5.1: *After Kleinschmidt & Jaeger (2015: 12). An example showing varying F2 locus cue distributions to 2 categories. The upper graphs show cue distributions for each category for three different speakers, and their respective posterior distributions as compared to a 'standard speaker', Norman (represented by the dotted line).*

This figure visualizes how a hypothetical listener's distributions of acoustic cues to membership in a phonological category (in this case /b/ or /d/) could develop for different speakers, and how the varying distributions might affect the listener's category judgments. For example, the distributions for Sherman overlap more extensively, leading to a more difficult categorization task as represented by the posterior distribution, while the very narrow cue distribution for Priscilla leads to a much smaller area of ambiguity. This reflects how a listener might make use of varying cue distributions to aid in their speech perception, and also how cues which overlap (as they do in the case of the model in this chapter) can lead to difficulty in making judgments.

One previous study that made use of such a model was Clayards et al. (2008), which investigated if manipulating the probability distributions of VOT (Lisker & Abramson, 1964)

would have an effect on listener categorization of sounds as /p/ or /b/ (i.e. voiced or unvoiced). The task given to the subjects was to discriminate between the words "Peach" and "Beach" by looking at pictures in an eye-tracking paradigm, and they were split into two groups: one group heard /p/ and /b/ phonemes where the VOTs were taken from probability distributions with a narrow variance of 8 ms, while another group heard tokens where the VOTs were taken from a distribution with a wider variance of 14 ms. Both distributions contained the same number of tokens, and the same category means. The hypothesis was that, if the theory behind the ideal listener model is correct, listeners would have less difficulty categorizing the phonemes when the cue distributions were narrower. The results of the study supported this hypothesis, with listeners' posterior distributions having much higher slopes (and therefore less category ambiguity) in cases where listeners received cues from the narrower distributions. There was, however, still ambiguity in both cases, illustrating the probabilistic nature of the categorization task.

Using this theory of the ideal listener, it will be possible to use the corpus data collected for the study in Chapter 3 to provide appropriate cue distributions that can be used to train a model to make a prediction regarding the formality of a given recording.

5.2 Model structure

5.2.1 Model specifications

At its core, the objective of the model is to be able to calculate a posterior probability of category membership for a given recording of spoken Japanese based on its mean f_0, articulation rate, and f_0 range, and then to make a specific prediction as to its category while taking into account the uncertainty due to the probabilistic nature of the decision. This can be done using Bayes' rule as described in Section 5.1.3, but it is first necessary to calculate both the likelihood and prior probability of category membership of a given token which will require the use of Bayesian inference. More specifically, this was done by using Bayesian generalized logistic mixed effects regression models using the *rstanarm* (Stan Development Team, 2016) package in R (R Core Team, 2017) on the entirety of the data set in order to calculate the priors, and generalized logistic models on the data subset by speaker to calculate the likelihoods. Arriving at the priors and likelihoods was a multi-step process that began with specifying the logistic models in *rstanarm* as in (5) and (6).

(5) Full Model: $y =$ Formality, $\beta =$ logf0 + lograte + lograng e,

$u =$ Intercept + Slope of logf0, lograte, and lograng e | Subject

Link Function: Logit

Data Used: Full training subset

Priors: Intercept: Normal – mean: -0.2, SD: 0.5

β: Normal – Mean: -4, SD: 5

(6) **Full Model**: y = Formality, β = logf0 + lograte + logrange,

Link Function: Logit

Data Used: Individual speaker training subset

Priors: **Intercept**: Normal – mean: -0.2, SD: 0.5

β: Normal – mean: -4, SD: 5

Data for use in these models was $\log_{10}$ transformed in order to put the variables on similar scales, which allows them to be combined together in the logistic regression as predictors with sensible weights (beta values). The model in (5) was used to calculate an intercept and slope for use in estimating the prior probabilities of category membership, while the model in (6) was used to calculate the likelihoods. The model in (5) can be read as *formality* as a function of *logf0*, *lograte*, and *logrange*, with random intercepts and slopes for each speaker. The logit link function means that the model will calculate the posterior value for each sample of the model as in (7).

(7) $p = \log(\frac{p}{1-p})$

The priors in the model specification refer to our expected prior values for the beta values of the model (i.e. the intercept and slopes) and critically *not* to our expected prior values of the fixed factors themselves.

The specification of priors is quite important to a Bayesian analysis, and so bears further discussion. The priors of the intercept in both models represents the probability of any given recording being formal, in log-odds space (i.e. after a logit link, as in (7)), and from the experiment

in Chapter 4 we have some information about listeners' prior expectations of the overall probability of formal versus informal speech in Japanese. Overall the subjects in the experiment appeared to expect informal speech to occur slightly more often than Formal (~55% of responses were informal to unmanipulated stimuli). Although this does not definitively show that all listeners expect informal speech to be more frequent than formal speech in Japanese, it will still be used to inform the priors. Given the expected ~45% chance of formal speech, the prior on the intercept will be a normal distribution centered on -.2 (the result of logit(.45)) with a fairly small amount of variance (DF = .5), but which nonetheless allows for the less-likely possibility of formal speech being more probable. The priors on the beta values are slightly more straightforward, as we know from the experiment in Chapter 3 that an increase in any of the three variables should indicate a higher likelihood of informal speech. The exact degree of the negative slope is not known, however, and so the prior on the beta values has a relatively high amount of variance (SD = 5), allowing for a range of possible slopes.

In order to further test the sensitivity of the posterior distribution to the prior specifications of the model, a series of test models were created using a number of possible priors. These models were tested on a training subset of the data, which is described in more detail in Section 5.3. The different sets of priors tested were as follows:

1) Informative priors. These are the priors shown in (5) and (6).

2) Uninformative priors. These are normal distributions centered on 0 with a SD of 10 for the β prior and 1 for the intercept prior.

3) Reversed priors. These priors suggest that we expect the opposite of what was found in the previous experiments. The β prior is a normal distribution with a mean of 5 and a SD of 5, and the intercept prior is a normal distribution with a mean of 0.5 and a SD of 0.5.

4) Exaggerated priors. These priors exaggerate the expected effects. These are normal distributions with a mean of -10 and SD of 5 for the β prior and a mean of -1 and SD of 0.5 for the intercept.

Calculating the posterior probabilities under these different sets of priors shows that the posterior distribution of this model is somewhat sensitive to the prior specification. Under these different priors, if a hypothetical recording had a roughly even chance of membership in either category under prior 1 (51.2%), then under prior 2 the same recording has a 45.9% chance of being formal. Prior 3 gives the same recording a 58.8% chance of being formal, while prior 4 gives only a 41.5% chance. This total gap of ~19% based on different priors (or ~8-10% difference from the informed priors) means that in somewhat ambiguous cases a change in priors is enough to reverse the predictions of the model. The differences in the model under these four different prior specifications is illustrated in Figure 5.2, which shows the posterior probability

densities of the model under each set of priors.

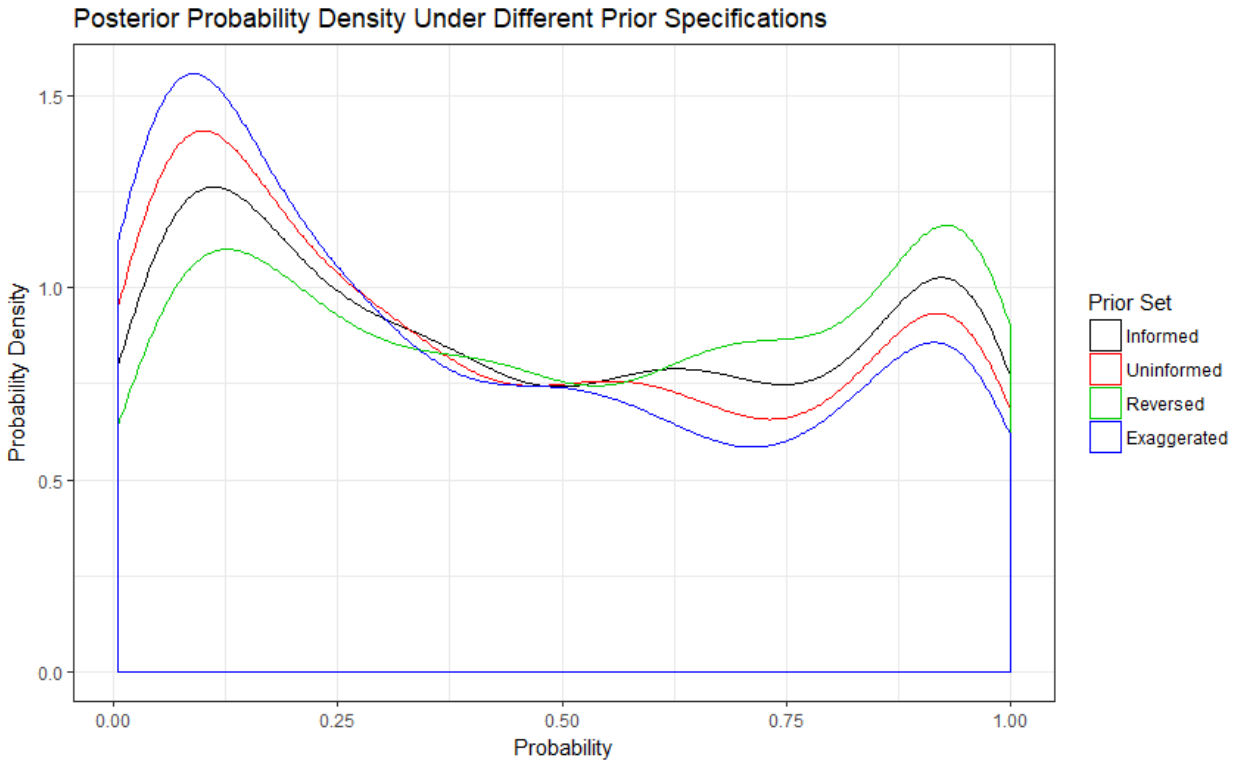


Figure 5.2: *Probability densities of the posterior distribution of the model under different prior specifications.*

Figure 5.2 shows values that correspond roughly to what we would expect based on the different sets of priors previously discussed. The informed priors expect more items to be informal than formal, but predicts recordings to be formal more often than all except for the reversed priors, which is the only model which appears to expect more recordings to be formal than informal. The uninformative priors expect more recordings to be informal, likely due to the slight skew in the training data towards informal speech, which together with the uninformative priors causing

the likelihood to dominate the posterior calculations results in more predictions of informal speech. The models using the reversed and exaggerated priors behave roughly as expected, pushing the posterior probability estimates towards either predictions of more formal or informal speech respectively. The salience of prior specification to the calculation of the posterior in the model demonstrates one of the concrete advantages of making use of the Bayesian framework for this model.

With the priors aside, the largest difference between the two models worth discussing is the random effects structure, which is not included in the model used to calculate likelihood. Not including any random effects in the model in (6) was a straightforward decision, as there are no reasonable random factors to consider – this model only looks at the data from a single speaker at a time and attempts to calculate the intercept and slopes for that speaker's data only, in order to represent a listener's distribution of cues specific to the speech perception situation at hand. On the other hand, the reasoning behind the decision to include random intercepts and slopes for each speaker in the model to calculate the priors is less straightforward. The logic behind including the random effect of speaker comes from considering what exactly the priors are supposed to represent in this model. Based on the previously discussed ideal listener model of speech perception, in this model the priors essentially represent the entirety of what a listener knows or believes about a cue distribution outside of the particular speech perception situation

they are in. Therefore, in order to have priors represent a listener's combined knowledge of cue distributions for a category, the intercept and beta values used to calculate the prior probabilities were the *average* of the random slopes and intercepts for all of the speakers, so that they would represent the combined sum of a listener's knowledge about cues to the category of formality across a number of situations.

5.2.2 Calculating posterior probability under uncertainty

The second task for the model was to use the intercepts and slopes from the models in (5) and (6) to first calculate likelihoods and priors, and then a posterior probability for each recording in the data set. Calculating these probabilities is fairly straightforward based on the results of a logistic regression, as an inverse logit function will calculate a probability from the log odds given by the regression. This can be done in R using the *plogis()* function, as shown in (8).

(8) $p = plogis(\beta_0 + \beta_1 x_1 + \ldots \beta_n x_n)$

Due to how the response data used to train the models is coded (1 = formal, 0 = informal) the result will be the probability of the given recording being formal.

The next step was to weight these probabilities in such a way that they better express the inherent uncertainty and probabilistic nature of the judgments being made. This was done by taking advantage of the posterior distributions for the beta values provided by the Bayesian models. In order to better express the uncertainty of the situation being modeled, a series of beta

values were taken from the 5%, 10%, and 15% credible intervals of the posterior (referring to the points on the distribution under which 5, 10, and 15% of the probability density lies, respectively) along with the mean (i.e. the most likely values) and were then weighted and summed as in (9) in order to arrive at a final weighted probability under uncertainty, where *p* alone refers to the most likely posterior probability, and p#a and p#b refer to the posterior values at the two points of the # percent credible interval.

(9) *.05+(p*.2 + p15a*.095 + p10a*.12 + p5a*.135 + p5b*.135 + p10b*.12 + p15b*.095)*

The weights for each of the probabilities was determined by their distance from the mean (most likely) value, here represented by *p*. In order to express an additional layer of uncertainty caused by the fact that the prosodic cues taken into account by the model are likely not all of the cues to the category of formality in Japanese, the scale is weighted in such a way that there will never be less than a 5% chance or greater than a 95% chance of category membership given. This process was then applied to all values in the data set to arrive at probability distributions for both the likelihoods and the priors. An example of such distributions for one of the speakers in the data set broken down by variable is shown in Figure 5.3. In Figure 5.3, we can see that for this particular speaker mean f_0 is not highly predictive of formality (i.e. no value of f_0 will lead to a probability of formality above 50%), while the other two cues do appear to be predictive, with articulation rate slightly more so. The models in (5) and (6) combine all of these fixed factors

into a single model, allowing their predictive significance to be weighted with appropriate beta values.

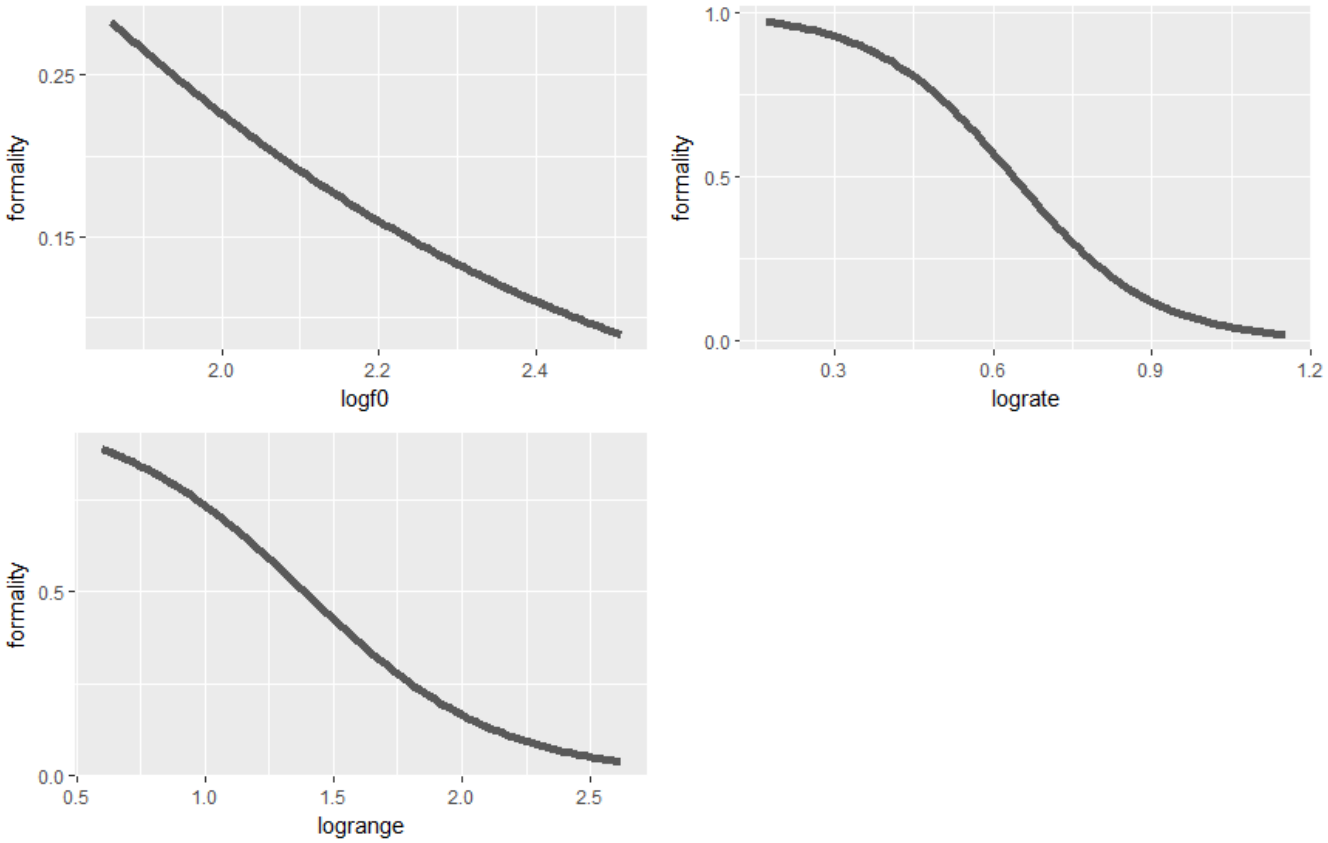


Figure 5.3: *Posterior probability distributions for the possible values of each prosodic cue for one speaker.*

Finally, once the likelihoods and priors had been calculated, Bayes' rule was used to calculate the posterior probability of formality for each recording in the data set. Predictions of the actual level of formality were based on this value, but with an added element of uncertainty. Guesses were made by creating a random normal distribution using the *rnorm()* function in R with a mean of the posterior probability, and a SD of .1, which was used to represent the level of uncertainty. Predictions were then made based on how much of the random normal distribution

was concentrated above or below 0.5 (the point at which category selection would be completely ambiguous). Making predictions in this fashion means that in the case of highly ambiguous recordings, there is a chance of some variation in the predictions of the model each time it is run. It also allows for the calculation of a level of uncertainty in the prediction, based on the percentage of the density of the random normal distribution that is concentrated on the side of 0.5 representative of the level of formality that the model predicted. For example, if the model predicted that a recording was informal, and 95% of the distribution was concentrated below formality = 0.5, there would be a low level of uncertainty in the prediction. However, if only 55% of the distribution was concentrated under 0.5, the response would be highly uncertain. This level of uncertainty was quantified, and included in the analyses described in Section 5.3.

5.3 Modeling results and analysis

Once predictions were made by the model, each prediction was labeled as either correct or incorrect in the data set. Two versions of the model were initially tested: one version which only took into account the prior probabilities calculated from the model in (5) while ignoring the likelihood values, and a second version which took the likelihood values into account. These models were trained on a subset of the data, in order to avoid biasing the likelihood with a heavily uneven distribution of formal and informal speech. Training the models on the full data set would

likely not produce good results due to the fact that the data set does not contain an even split of formal and informal utterances – in fact, ~80% of the data is informal. This means that any models trained on the full data set would be calculating the likelihood based on the priors of informal and formal speech being ~.8 and ~.2 respectively, which will lead to the great majority of the recordings being classified as informal if there is any degree of uncertainty. This effect can be overridden to an extent by the specification of informative priors (as shown in Figure 5.2), but the likelihood component of the model would still be heavily biased towards informal speech.

This potential problem of over-fitting was addressed by training the models in (5) and (6) on semi-random subsets of the data. These subsets were made up of random selections of recordings from each speaker which always included as many formal recordings as possible up to a maximum of 50 (though only two speakers had more than 50 formal recordings in their data), and 50 randomly selected informal recordings. This resulted in a subset of the data with much closer to a 50/50 split of formal and informal recordings, although there were still more informal recordings than formal as not all speakers had at least 50 formal recordings in their data set (in total 402 formal and 500 informal were included in the subset). In spite of this remaining imbalance the size of the subset was not reduced as doing so could cause the slopes of the models to conversely be underestimated due to a lack of data. In total the models were trained on five different random subsets of the data, although in practice this meant five different random groups

of informal recordings due to the lack of speakers with more than 50 formal recordings in their data. The intercepts and slopes of the models run on these different subsets were averaged, and the posteriors were calculated from the main data set based on the results.

The comparison between the two models was done as a test of the theory that making use of an ideal listener model (as discussed in Section 5.1.4), which includes likelihood information, will allow the model to achieve greater predictive accuracy. The results of running these two models were that the priors-only model was able to guess formal speech accurately 60.2% of the time, and informal speech 67.4% of the time (overall 67.4% accuracy). When the likelihood values from the individual speaker models were included, accuracy increased somewhat to 63.8% for formal speech, and 74.3% for informal speech (overall 72.5% accuracy). The accuracy of these models is significantly greater than chance for both categories of speech, although both are better at categorizing informal speech, which likely indicates that the imbalance of formal and informal speech in the training data leads the model to guess that speech is informal more frequently. This imbalance in accuracy categorizing the different types of speech is mostly due to the skewed likelihood in the training data, combined with a prior on the intercept which tells the model to expect informal speech slightly more often. To test this intuition, a third model was run using the 3rd set of priors discussed in Section 5.2.1 which suggested that the model should expect formal speech more often than informal. Using these 'reversed' priors had the expected

effect: overall accuracy was decreased slightly (to ~68.5%), but the issue of the model being better at predicting informal speech largely disappears (and in fact reverses), with the model accurately categorizing informal speech ~67.9% of the time and formal speech ~71% of the time. A comparison of the results of the three model types can be seen in Figure 5.4.

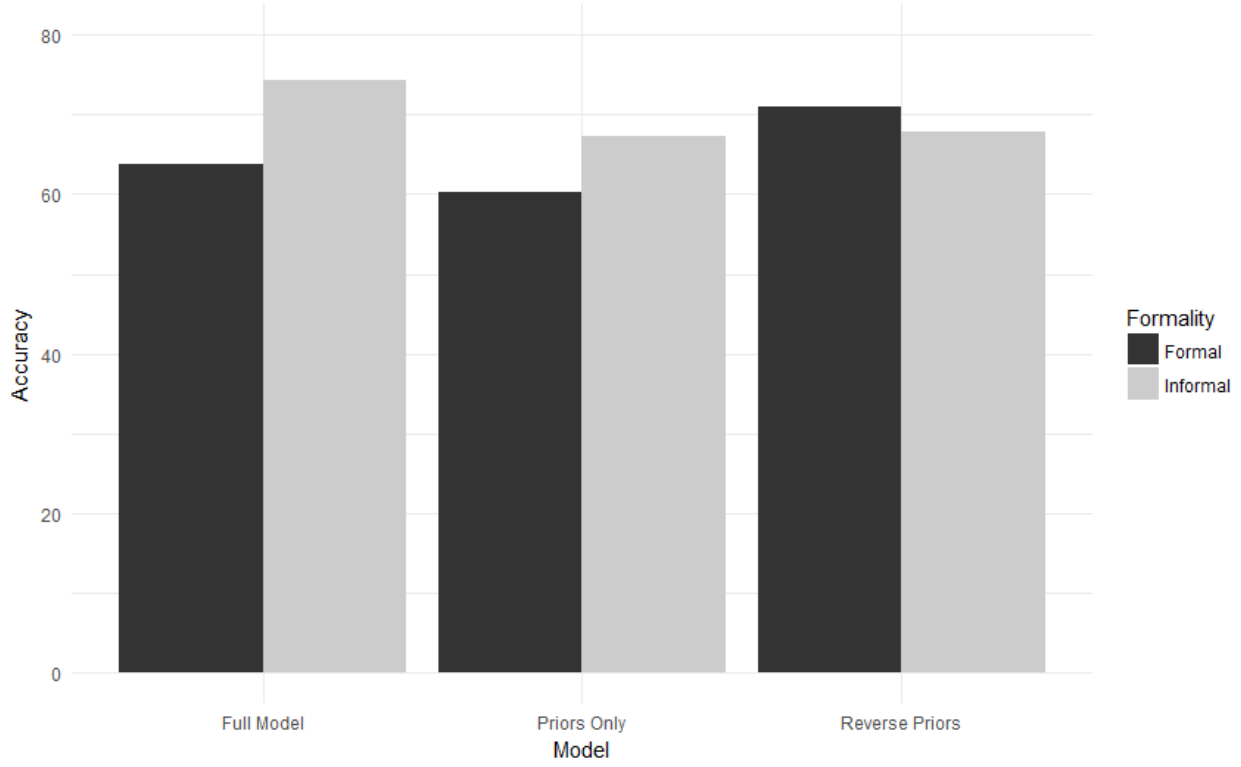


Figure 5.4: *Comparison of the categorization accuracy of the three model types tested by actual level of formality of the recording.*

An additional point of interest is that there is a significant statistical relationship between the model's predictive accuracy and the level of uncertainty (as described in Section 5.2.2). Based on a simple mixed effects model with uncertainty as the dependent variable, accuracy as the fixed factor, and the speaker of the recording as the random factor uncertainty is ~12% higher (the

fixed effect slope estimate) in cases where the model guessed incorrectly. This is not particularly surprising by itself, but it does indicate that the accuracy of the model could potentially be improved by finding ways to decrease the ambiguity of certain recordings, perhaps by adding more significant predictors to the model.

5.4 Discussion

The level of accuracy of this probabilistic model in predicting the level of formality of recordings based on their prosody has a number of potential implications. Firstly, it lends some qualified statistical support to the theory discussed in Chapter 4 that listeners may refer to probabilistic distributions of phonetic parameters as cues to use in their judgments of category membership. When using the *predict()* function in R to attempt to guess formality using only a logistic regression that does not consider likelihood, or the inherent uncertainty of the judgment being made (as in the logistic regression analysis described in Chapter 3), there is only a ~6% chance of accurately predicting formal speech. A linear discriminant analysis (Fisher, 1936) performs much better, and is able to identify formal speech accurately ~59% of the time, although it remains better at categorizing informal speech than formal regardless of what prior probabilities are given for the two categories. While the Bayesian model is not massively more accurate than a discriminant analysis, its additional sensitivity to the prior distributions allows for it to be more

balanced in its predictions for both categories. The Bayesian model's computational process of referring to cue distributions for the prior and likelihood appears to be a superior method for modelling listener categorization judgments if the prior specifications are properly considered.

In Japanese specifically, the results of the model lend support to the overall hypothesis of this thesis that it is possible to model and predict formality in speech without making use of any lexical, morphological, or phonological information. More generally, the success in using prosody to predict formality could indicate that it may be possible to modify the probabilistic model to predict other binary linguistic (such as e.g. phoneme classification) or meta-linguistic (such as e.g. the presence of irony) categories as well, given some prior knowledge about the relationships between prosody and the categories of interest. The model described in this chapter could therefore be a more general tool for modeling probabilistic phenomena in Japanese (or other languages).

In terms of experimental insight, as discussed in section 5.1.2, if we assume that the model presented in this chapter is a reasonable parallel for the cognitive process of category judgment in human listeners then the results of this model can be used to inform baseline expectations for categorization accuracy under fairly ideal conditions. Given that this model hovered around 70% accuracy, if a human listener performs noticeably better or worse (as was the case in the experiment in Chapter 4) then it is possible that there is some other (possibly unconsidered) factor

that is influencing the results. In the case of the experiment in Chapter 4, the fact that the version of the model that included likelihood information was more accurate than the one which did not might lead us to conclude that the lack of this information caused a similar (or greater) drop in accuracy for the human listeners. The listeners were only played a single recording of each speaker, which was likely not enough to establish the accurate likelihood distributions that the model has access to. If, conversely, the subjects had been more accurate than the model we might conclude that there was some aspect of the experimental design which was giving the listeners additional, perhaps unintended information.

Although the model can discriminate between formal and informal speech fairly accurately, there are still some ways in which it could be improved. One way in which the accuracy of the model could potentially be improved is simply by having a larger and more balanced training data set. A much larger corpus of data which would allow an even split of formal and informal utterances in the data used to train the model without sacrificing sample size would be ideal, potentially allowing for greater overall accuracy. Additionally, it is very unlikely that the three prosodic variables (mean f_0, articulation rate, and f_0 range) are the only acoustic cues to formality present in Japanese. There are many potential cues, such as e.g. intensity, vowel quality, f_0 slope, and pause frequency, which were not investigated in this thesis but could nonetheless have some relationship with formality in Japanese. Given that incorrect predictions

by the model were characterized by increased uncertainty, including further fixed factors in the model could potentially decrease the uncertainty of predictions (although likely not ever eliminate it completely) and lead to increased predictive accuracy, although it does risk model over-fitting.

Chapter 6

Discussion

6.1 Summary of results

This thesis investigated the relationship between prosody and formality in spoken Japanese. The experiments and statistical tests described in Chapters 2 through 5 – the results of which will be reviewed in the remainder of this section – presented evidence that there is a significant relationship between the prosody of an utterance and its level of formality, from both a speech production (Chapters 2 and 3) and perception (Chapter 4) perspective. This final chapter will first review and summarize the results of each experiment discussed in the main body of the thesis, as well as the Bayesian statistical model described in Chapter 5. These results will then be discussed as they relate to the initial research questions and hypotheses of this thesis as put forth in Chapter 1, as well as discussing how the results relate to past work on the subject, and their implications for future speech research. The methodology used throughout this thesis – both experimental and statistical – will also be discussed as it relates to refining experimental approaches to the study of politeness and formality. Finally, potential future research directions related to the topic of this thesis will be discussed, followed by concluding thoughts on the overall findings presented in the preceding chapters.

6.1.1 Formality and prosody in speech production

Chapters 2 and 3 of the thesis investigated the relationship between prosody and formality in Japanese from a speech production perspective. Chapter 2 was a smaller pilot study involving

controlled stimuli produced in the lab, while Chapter 3 was a larger, corpus-based study. The pilot study in Chapter 2 aimed to discover significant relationships between the prosodic variables of mean f_0, duration, and amplitude and formality in spoken Japanese in order to determine what variables should be investigated further in the corpus based study in Chapter 3. Analysis of recordings collected in the lab using mixed effects regression models showed a significant relationship between only mean f_0 and duration and formality, with no significant relationship between amplitude and formality. Following these results, along with manual observation of the data, and some evidence from studies of other languages (Winter & Grawunder, 2012), mean f_0, articulation rate, and f_0 range were chosen as the variables to be studied in the corpus based study.

Chapter 3 analyzed a 10-hour corpus of recordings of conversational Japanese created specifically for use in this thesis. The recordings from each speaker were labeled and segmented, and f_0 and articulation rate were measured. As there were a number of pitch-peak estimation errors present in the data, such errors were diagnosed and corrected via the use of an automated MATLAB script (discussed further Section 3.4, and additionally in Section 6.3 below). Analysis using mixed effects regression models showed a significant relationship between formality and all the prosodic variables investigated – mean f_0, articulation rate, and f_0 range – where each variable was significantly higher in informal recordings than in formal recordings. In addition, a functional data analysis (Ramsay, 2006) was conducted in order to further analyze f_0 range, where

polynomial functions were fitted to the f_0 vectors of each utterance, and the orthogonalized coefficients of the functions were related to linguistic variables (Grabe et al., 2007). This functional data analysis showed that f_0 in informal speech falls lower and peaks higher than in formal speech, and also that the quadratic term of the polynomial (i.e. breadth of curvature) was significantly lower (i.e. a narrower curve) in informal speech. Further details of this analysis can be found in Section 3.6.

6.1.2 Formality and prosody in speech perception

Chapter 4 investigated the relationship between prosody and the perception of formality via an experimental task where subjects were asked to categorize de-lexicalized speech recordings as formal or informal given only prosodic information. The de-lexicalized speech stimuli were synthesized based on the natural speech recordings used in the study in Chapter 3, and the same prosodic variables investigated in Chapter 3 were then manipulated upwards (*ex hypothesi*, to sound more informal) or downwards (to sound more formal). The relationship between the difference in subjects' judgments between a prosodically manipulated recording's formality and a 'base' unmanipulated recording and the direction of the manipulation (upwards or downwards) was then analyzed using mixed effects regression models.

The results of the analysis showed that there is a relationship between changes in the prosody of a recording and a listener's judgment of those recordings as formal or informal, but it

is not completely straightforward. When individual variables were manipulated, only mean f_0 showed a significant relationship to the direction of prosodic manipulation, wherein listeners judged recordings as more informal when mean f_0 was manipulated upwards, and more formal when it was manipulated downwards. However, this pattern did not hold with the other variables, and listeners were actually slightly more likely to judge recordings as formal even when the variables were manipulated upwards. However, when all three variables were manipulated together, the pattern was much clearer, showing a significant relationship where there was a difference of ~.7 of a step on the Likert scale based on the direction of manipulation. Further details regarding this relationship can be found in Section 4.3.3.

The hypothesis proposed to explain these results was that when listeners make category judgments – such as between formal and informal – they access a probabilistic distribution of acoustic cues to these categories based both on their overall prior knowledge of how these cues are likely to be distributed, and on how the listeners expect them to be distributed in the specific speech context (see e.g. Clayards et al., 2008; Norris & McQueen, 2008; Kleinschmidt & Jaeger, 2015). Under such a theory, given that the distributions for any single cue to the formal and informal categories will overlap quite a bit according to the data observed in Chapter 3, and that the distribution for formal speech is narrower than that of informal speech, it makes sense that listeners have a harder time categorizing speech as informal unless the cue quite clearly falls into

the informal distribution (Figure 4.4 in Chapter 4 illustrates this theory). This theory was used as a basis for constructing a probabilistic Bayesian model of formality in Japanese.

6.1.3 Probabilistic modeling of formality in Japanese

Chapter 5 described a Bayesian statistical model of perceived formality in spoken Japanese which used probability distributions of prosodic variables and prior knowledge about the relationship of each prosodic variable with formality to predict the formality of recordings of spoken Japanese. This was accomplished via the analysis of Bayesian generalized logistic regressions using the *rstanarm* package (Stan Development Team, 2016) in R, while also accounting for the inherent uncertainty of the category judgment task. Further details of the structure of this model can be found in Section 5.2.

The model aimed to replicate the hypothesized cognitive process underlying the category judgment task, with the goal of obtaining a baseline expectation for categorization accuracy which could be used to further explain or evaluate the performance of the experimental subjects in Chapter 4. The results were that, when trained on a random subset and given reasonably informative priors, the model was able to accurately predict the level of formality of the recordings in the data at an overall rate of 72.5%, though formal speech was identified correctly less often (63.8%) than informal speech (74.3%). This issue of the model over-categorizing speech as informal could be overcome by providing the model with priors that instructed it to

expect more informal speech, which demonstrated the model's sensitivity to prior specification.

6.2 Discussion of results

6.2.1 Research questions and hypotheses

Chapter 1 of this thesis set out a number of research questions to be addressed and corresponding hypotheses to be tested in the studies described in Chapter 2 through 5. Table 6.1 reviews these research questions and hypotheses.

Table 6.1: *Research questions and hypotheses investigated in this thesis.*

Research Question	Hypothesis
1) Do speakers of Japanese make use of changes in prosody to help signal their intended level of formality in conversation?	*Speakers do use changes in prosody to express different levels of formality in speech.*
2) What specific *differences in the prosody of an utterance are used by speakers to distinguish different levels of formality?*	*Differences in f_0, articulation rate, and f_0 range will co-vary significantly with the level of formality in speech. Each of these variables will be higher in informal speech.*
3) Can listeners make use of prosodic cues to help determine the intended level of formality of the speaker?	*Changes in the prosody of an utterance will have a relationship to listeners' category judgements of speech as formal or informal.*
4) Given the results of the experiments investigating questions (1)-(3), is it possible to build an accurate predictive statistical model of formality in Japanese?	*Using the information from the experiments investigating hypotheses (1) – (3), a predictive Bayesian probabilistic statistical model will be able to accurately model the relationship between prosody and formality in spoken Japanese.*

Questions (1) and (2) were investigated by both the pilot study in Chapter 2, and the corpus based

study in Chapter 3. The results of the studies in these chapters showed largely unqualified support for hypotheses (1) and (2) – although it is not clear whether or not speakers were making use of active knowledge of the prosodic properties of formality in their speech production, the significant relationships between prosody and level of formality found in the corpus based study indicate that prosody does play a role in the expression of such paralinguistic information, in spite of the fact that it is reliably indexed by various lexical and grammatical items (Cook, 1998; Okamoto, 1999). These effects occurred across all speakers (see Appendix II for figures), with the coefficients of mixed effects models showing the phonetic patterns of formal and informal speech to be consistent regardless of age, or gender. Similarly, the results showed a pattern where each prosodic variable examined was significantly higher in informal speech as compared to formal speech. The support for hypothesis (2), while consistent, is nonetheless variable. There was variation in the magnitude of the main effects of the different variables among the different speakers, with for example one speaker showing an average difference of only ~3 Hz between formal and informal speech, while another speaker showed a difference of ~25 Hz. Although the same significant pattern was observed consistently in each speaker's data, the variation does indicate that different speakers might use somewhat different strategies in their use of prosody in the expression of different levels of formality. In other words, although the pattern is consistent and significant on average, that does not mean that it is identical for all different speakers, and

speech contexts. The type of speech used in the corpus-based study was fairly homogeneous – all speech from Tokyo area speakers, collected conversationally in an interview format – and there was variation even within that data set. It is possible that different patterns would emerge in different varieties of speech. The implications of this variation will be discussed further in Section 6.2.3.

The results of Chapter 4 do appear to support hypothesis (3), although the results were less straightforward than those of Chapters 2 and 3. The results seem to support the hypothesis that prosody is used by listeners to make category judgments of formality, and that these judgements are made probabilistically. As discussed in Section 6.1.2, listeners appeared to have difficulty categorizing recordings consistently when only a single variable was manipulated, indicating both that they may be accessing overlapping cue distributions to make category judgements, and also that manipulating only one variable is not sufficient to decrease the ambiguity of the *overall* distribution of prosodic cues to category membership. Although at first glance the fact that manipulations of individual prosodic variables do not appear to have much effect on listener's category judgments would appear to weaken the support for hypothesis (3), the fact that manipulating all three variables conversely has a significant effect on their judgements points to a different conclusion. This result indicates that the effects of each individual variable on the categorization task is *super-additive*, meaning that listeners are consistently

making use of *all* available information in order to make uncertain predictions. Changing a single variable may only increase the ambiguity due to the inconsistent information being received by the listener, but if enough cues are manipulated together, to fall into the distributions that a listener expects for a certain category, then the change in their predictions of category membership becomes more pronounced and consistent.

Taken together, the results of the production studies – which showed consistent relationships, but with variation amongst speakers – and the perception study – which suggested that speakers make use of a number of prosodic cues to category membership taken as a group in a probabilistic fashion – led to the construction of a probabilistic Bayesian model which was able to predict the category membership of recordings with an accuracy ~20% better than chance, supporting hypothesis (4). Although the model is not perfect, it achieves its level of accuracy by taking both individual speaker variation and listener uncertainty into account when modeling formality. It is not terribly surprising that the model was not able to attain an accuracy rate anywhere approaching 100%, considering that human listeners had a difficult time making the formal/informal category judgments in the experiment in Chapter 4 when only un-manipulated prosodic information was available. In such cases there was no significant relationship between the recording's formality and listeners' judgments, indicating that they did not perform significantly better than chance. Given that the model has more complete information about the

cue distributions of the individual speakers than is available to human subjects, it is logical that the model would perform slightly better than the human speakers if it is in fact true that they rely on cue distributions specific to the speech context in order to make predictions of category membership. The model demonstrated this intuition empirically, as a model which did not consider information from the distributions of individual speakers had a lower overall rate of accuracy (66.2% vs. 72.5%).

6.2.2 Prosody and formality

The results of the studies in Chapters 3 and 4 have produced a number of insights regarding the relationship between prosody and formality in both speech production and perception, as they relate to Japanese and potentially to other languages as well.

6.2.2.1 Relationship between prosody and formality in speech production

To recap, the studies in Chapters 2 and 3 showed that in Japanese conversational speech the mean and range of f_0, as well as articulation rate are all significantly higher in informal speech. This result contrasted with much of the previous literature on the relationship between prosody and formality in Japanese – which had found either mixed results when analyzing f_0 (Ofuka et al., 2000; Ito, 2002) or that f_0 was higher in polite (formal) speech (Loveday, 1981; Ohara, 2001; 2004; Tsuji, 2004) – and also going against the predictions of the *frequency code* (Ohala, 1984) that higher f_0 should be a universal linguistic property of polite speech. Conversely, the results

largely mirrored those of cross-linguistic studies of Korean (Winter & Grawunder, 2012) and a recent study of Catalan Spanish (Hübscher et al., 2017), indicating some potential cross-linguistic tendencies (if not necessarily universals) in how phonetic parameters such as f_0 or intensity are used in the expression of formality.

The contrast of the current study with the results of previous related studies of Japanese are somewhat surprising, but can likely be explained by methodological differences. While the current study judged formality post-hoc based on the presence of lexical items or grammatical structures indexical of different levels of formality, previous work judged utterance formality based on either the relative social status of whomever the speaker was addressing (in the case of Ito, 2002) or instructed subjects to role-play as if they were addressing a status superior or inferior (Ofuka et al, 2000; Tsuji, 2004). Loveday (1981) is in some ways most similar to the current study in that it evaluated politeness based on the inclusion of certain 'politeness formulae', or specific words or phrases judged to be particularly polite, although the speech was read rather than spontaneous, and there are known differences between the phonetic parameters of read and spontaneous speech in Japanese (Nakamura et al., 2007). These methodological differences make it difficult to directly compare the current study to previous work on the subject in Japanese, and although the differences in results do cast doubt on the validity of some previous findings, they do not necessarily invalidate them as it is possible that prosody relates to formality differently in

different types of speech (i.e. read vs. elicited vs. conversational).

Another point of interest is the potential cross-linguistic validity of the results of this thesis. The results of this study concerning f_0 measures were extremely similar to those of Winter & Grawunder (2012) for Korean, and Hübscher et al (2017) for Catalan Spanish. Those two studies, along with the current study, contribute to a growing body of evidence against previous claims that high pitch correlating with polite or formal speech is a linguistic universal (Ohala 1984; Brown & Levinson 1987; Gussenhoven, 2002). Hübscher et al. (2017) in particular notes that the similarities in how prosody is used to encode formality may simply be ascribed to similarities between the cultural interpretations of formality and politeness, noting a case where speakers of German from different cultural contexts (Germany and Austria) use prosody to express paralinguistic information in different ways (Grawunder, Oertel & Schwartze, 2014). Interestingly however, Japanese is also posited as a contrastive example, where high pitch may be interpreted as a sign of submissiveness (Hübscher et al, 2017: 155) meaning that the fact that this thesis also found higher f_0 in informal speech suggests that the issue may be more multi-faceted. This does not invalidate Hübscher et al (2017)'s observation, but it does show that the explanation may be somewhat reductive. These cultural interpretations may not be entirely static, and variation in speech context (conversational speech, in the case of Chapter 3) may produce unexpected variation in results where prosody is concerned.

In addition to the previously discussed measures of f_0, Chapter 3 also contained a functional data analysis of f_0 at the approximate level of the accentual phrase. While the analysis of this study was focused primarily on the suprasegmental level (i.e. overall differences in prosody rather than differences at the level of individual segments) which makes the results difficult to interpret in terms of the structure of Japanese, the functional data analysis does allow us to see some ways in which the results are connected to the prosodic structure at the accentual phrase level. As was previously noted, Tokyo Japanese is characterized by an accentual phrase structure where there is an initial lowering of f_0, followed by a rise to a pitch accent, and finishing with a (possible) final drop in f_0 (Pierrehumbert & Beckman, 1988). The results of the functional data analysis indicated essentially that the properties of the accentual phrase were exacerbated in informal speech; there was a lower initial start, a higher rise, and a deeper final drop in f_0, which indicated an overall increase in *pitch dynamism* (Henton, 1995) in informal speech in Japanese. While this study is limited to showing this difference at the level of the accentual phrase, it is reasonable to expect that similar results of increased pitch dynamism could be seen in other structural contexts, such as for example increased phrase-final f_0 movement, as was seen in informal speech in Ofuka et al (2000). It would be useful in future research to investigate similar phenomena in further distinct contexts, which would allow the results to be more easily interpreted in terms of the specific structure of Japanese.

6.2.2.2 Prosody and formality in speech perception

The study in Chapter 4 showed that there is a connection between prosody and the perception of formality in Japanese wherein changes in mean f_0, articulation rate, and f_0 range will influence listener category judgments. The further critical observation from the results was that the manipulation of any individual phonetic parameter was not enough to influence listener judgments, but that the effect of the manipulation of multiple variables was super-additive, and could produce the observed effect. This result can help explain some of the differences between the results seen in this thesis and those of previous studies of the perception of formality in Japanese (Ito, 2001). The main result from Ito (2001) that the results of this thesis contradict was that articulation rate did not appear to be used consistently by listeners in the judgment of politeness. This inconsistency can be explained by the differences in methodology between the two studies – as Ito (2001) considered each prosodic variable individually, and did not control the differences in the variables in the different stimuli, it is possible that utterances containing a higher speech rate also contained lower mean f_0, causing ambiguity in listener category judgments when all the variables were considered together.

While the results in this thesis differed from Ito (2001), they patterned quite closely with those of Laan (1997), which also sought to measure the effect of changes in prosody on a category judgment task, in this case listener categorization of speech as either read or spontaneous. Laan

(1997) found that manipulation of individual prosodic variables did not reliably shift a listener's categorization from one category to another, but when all of the prosodic variables were manipulated together, listeners would consistently choose the opposite category. This was taken by Laan as evidence that no individual cue could be used consistently by listeners to distinguish between read and spontaneous speech, but rather that a complex and perhaps variable set of cues was being used. Based on the theory espoused in Chapter 4, it would be more accurate to say that the manipulation of individual cues is not sufficient to shift a listener's judgment, although it could be enough to increase ambiguity. It does appear to be the case that the manipulation of individual prosodic variables in Laan (1997) did increase the uncertainty of the category judgment, as listeners categorized the different types of speech significantly less consistently when any variable was manipulated as compared to an un-manipulated control condition.

In terms of this theory of the cognitive processes underlying the category judgment task, the results in Chapter 4 do suggest some qualified support, although based solely on what was observed in that study the hypothesis remains speculative. It is tempting to adopt such a theory based on the success of similar theoretical approaches taken in the fields of automatic speech recognition (Baker, 1975, Bahl et al., 1983; Goel & Byrne, 2000) and signal processing (see e.g. Kay, 1993; Fitzgerald & Rayner, 2000; Wang et al., 2002), but although it does provide a reasonable explanation for the results of Chapter 4, those same results also leave open some

questions which would require further research to address. More specifically, the super-additive nature of the results begs the question of whether, under such a theory, listeners are accessing distributions of individual cues, or whether they consider all of the potential prosodic factors in a holistic fashion, perhaps accessing a more abstract combined distribution. This question could be addressed in future research with a study similar to that described in Chapter 4, but with much closer control of the 'base' values of the phonetic parameters in question, and more variation in the degree of the manipulation of individual variables. This variation in the degree of manipulation would allow the theory that a large change in an individual cue (larger than those made in Chapter 4) could move that cue outside of the range of any great amount of ambiguity. If the purely super-additive results from Chapter 4 were to persist in such a case, it would be evidence against a theory under which listeners access and consider cue distributions individually.

6.2.3 Implications for future research

Based on the results of this thesis, there are some points that should be taken into consideration in similar future research. First is the fact that there appears to be a notable divergence between the results of previous studies of the relationship of prosody and formality in Japanese (as discussed in Section 6.2.2) and the results of this thesis. Some possible reasons for these differences have been discussed in Section 6.2.2.1 and Section 6.2.2.2, but more broadly it appears that it will be important not to generalize previous results regarding the relationship between

prosody and paralinguistic information (such as formality) from one speech context to others. This thesis has demonstrated that it is possible to obtain an opposite result when investigating such a relationship when a different type of speech is analyzed (conversational speech, in this case) or if different methodologies are applied (discussed further in Section 6.3). As the field of linguistics moves towards making greater use of corpora of natural speech, it will be useful to anticipate that some previous results may not be reproduced in an analysis of more natural speech, as was the case for previous results that higher f_0 indicates more polite speech in Japanese.

The implications from a speech perception perspective remain somewhat more speculative, but as discussed in Section 6.2.2.2 a main point of importance is that it appears that studies which relate phonetic parameters (such as e.g. f_0, articulation rate, and possibly others) must be cautious of under-interpreting their results and must carefully consider all the relevant parameters taken together rather than focusing on particular cues individually. The question of exactly how broadly applicable this observation is to speech perception remains open, but the similarity of the results of this thesis to those of Laan (1997) do support the possibility that other perception tasks involving multiple cues could see similar results.

6.3 Discussion of methodology

6.3.1 Methodology for studying meta-linguistic information in speech

As was mentioned in Section 6.2.2.1, the studies described in Chapters 2 and 3 of this thesis took a methodological approach to the analysis of paralinguistic categories that differed from those that had been employed in previous similar studies (Loveday, 1981; Ofuka et al., 2000; Ito, 2002; Tsuji, 2004) in a number of ways. In previous work, the paralinguistic category of 'politeness' was controlled either by attempting to directly instruct speakers to produce alternatingly polite or informal utterances (Loveday, 1981, Ofuka et al., 2000) or by recording the conversations of speakers of different social statuses (Ito, 2002; Tsuji, 2004). Although in the case where subjects were explicitly instructed to produce a certain level of formality the different results found in this thesis could potentially be explained by the difference in the type of speech being analyzed (read vs. conversational speech), in the latter case the types of speech analyzed in Ito (2002) and the current study were quite similar. Additionally, the argument that the difference in speech types causes the difference in results in the cases of Loveday (1981) and Ofuka et al. (2000) is weakened by the fact that the speech analyzed in Chapter 2 (which was read from written prompts) also exhibited a similar pattern to the data in Chapter 3. Because of this, it appears that the methodology used when determining whether speech is formal or informal is of importance, in addition to any differences that may be caused by the types of speech being analyzed.

In the pilot study in Chapter 2, rather than explicitly instructing speakers to produce formal or informal speech, the desired informal register of speech was elicited indirectly via the inclusion of word forms that were indexical of an informal register of speech – geminate contractions (see Chapter 2 for more details). Additionally, the carrier sentences containing those indexical items were semantically sensible and natural, with the hope that such sentences would help elicit a more natural style of speech. Given that the relationship between prosody and formality observed in Chapter 2 was very similar to that shown by the analyses of conversational speech in Chapter 3, it appears that, at least in Japanese, making use of indexical forms to elicit desired speech registers is effective in eliciting paralinguistic categories. Given that the results contrast with studies where different levels of formality were elicited directly, it is possible that the indirect method is better than explicitly instructing speakers to produce certain categories. It might be more difficult in languages other than Japanese where formality is not so explicitly indexed by certain words and grammatical forms to elicit specific registers in this way, but it is nonetheless worth noting that attempting to do so without explicitly informing the subject of what type of speech is being targeted may result in more appropriate data.

In Chapter 3, the primary difference in methodology as compared to Ito (2002) was that judgments of formality were made post-hoc, rather than assuming that levels of formality would remain consistent based on the relative social standings of the participants in the interview. This

approach, while more onerous for the researcher, allows for much more accurate evaluation of the level of formality of each utterance, and also allows potential switching between different speech registers over short periods of time to be taken into account. Once again, this methodology is more well-suited to analyzing Japanese than languages with less explicitly indexed levels of formality, but it does appear that not attempting to do so (as in Ito, 2002) may result in false negatives, or misleading results.

6.3.2 Methodology for the study of prosody

This thesis made use of a few methodological techniques which are of broader relevance to the study of prosody cross-linguistically. Firstly, the study in Chapter 3 made use of an automated MATLAB script (described in Section 3.4) which was able to diagnose and correct pitch-peak estimation errors in the f_0 vectors calculated from the corpus recordings. This was critical to the study, particularly due to the fact that f_0 range was one of the variables being examined, and having doubled pitch peaks would have drastically altered any analysis. These pitch-doubling errors are an issue for all current pitch tracking algorithms, but they are sometimes ignored in studies which analyze f_0 (as was the case with Winter & Grawunder, 2012). This is particularly problematic for studies which seek to investigate effects related to pitch range, where values are often acquired by taking the difference between the 5th and 95th percentiles of the f_0 vector (Winter & Grawunder, 2012), or by examining the standard deviation (as was the case in Chapter 3). All

in all, it is important for any study concerned with analyzing f_0 to make a concerted effort to eliminate these pitch peak estimation errors.

Related to the topic of f_0 range, the study in Chapter 3 also made use of a functional data analysis (Ramsay, 2006) in order to investigate the properties of the f_0 vectors from each level of formality in greater detail. This methodology created polynomial functions fitted to the f_0 vectors, and allows for direct comparison between the linguistic properties of the vectors, including the speed of change in f_0, the overall shape of the vector, how sharply f_0 rises and falls, and the mean f_0. This methodology allows for both visual and statistical comparison of these aspects of different groups of f_0 vectors (as shown in Figure 3.8 in Chapter 3) and is a valuable addition to any analysis of f_0.

On the speech perception side, while the methodologies used were not entirely novel, the contrast with the null results from a similar study (Ito, 2001) does demonstrate their effectiveness when analyzing the relationship between prosody and speech perception. In the experiment in Chapter 4, subjects were given information on the speakers' genders and ages, as well as being played short clips of each speaker’s speech before the experiment began, theoretically allowing them to access more specific prior information than they would be able to if they were entering the experiment with no expectations at all. Ideally such additional information would allow the speaker to make use of prosodic cues in a manner closer to how they would in a natural speech

context – i.e. making use of both their general knowledge about the category in question, and their specific expectations about the speaker they are listening to.

6.3.3 Statistical methodology

The statistical analyses in Chapters 2 through 4 made use of mixed effects regression models and cumulative link mixed models, rather than the more commonly used t-test or ANOVA. The reason behind this decision was that none of the data analyzed in this thesis satisfied the *independence* criterion of linear models, meaning that there were always multiple observations from each subject, resulting in within-group variation that an ANOVA cannot account for. The value of this approach was demonstrated in Chapter 2 – when analyzing the relationship between formality and utterance duration, it appeared (see Figure 2.3 in Chapter 2) that differences in utterance duration might be a main effect of the presence or absence of gemination, rather than of the formal/informal contrast. However, when examining the coefficients of a mixed model with random slopes for each subject, it was apparent that the relationship with the geminate/singleton contrast was not consistent among the speakers (with 2 of 5 subjects showing an opposite effect), while it *was* consistent among all subjects based on the formal and informal categories. These slopes can be seen in Appendix I. An analysis using only ANOVA shows both relationships as significant, and would have thrown the usefulness of duration (or articulation rate) as a cue to the formal or informal categories into doubt.

Chapter 5 of this thesis made use of Bayesian Inference, rather than frequentist statistical methods, when constructing a predictive model of formality in Japanese. The Bayesian framework was used because it can make use of multiple distributions of acoustic cues to the categories being predicted, in line with the probabilistic theory of the categorization task put forth in Chapter 4. This was ultimately an effective approach to modeling the relationship between prosody and formality in Japanese, particularly when compared to an attempt using the frequentist approach – attempts to discriminate between categories using a mixed effects logistic regression were accurate only ~6% of the time, as compared with ~63% of the time (or ~71% of the time in the reversed priors model). A linear discriminate analysis was more effective, with a ~59% accuracy rate when attempting to discriminate formal speech, but it still fell short of the Bayesian model. This comparison shows that a properly considered Bayesian approach using the appropriate priors is likely a superior method for modelling categorization tasks.

6.4 Future research directions

This thesis has produced many significant findings, and some of them raise further research questions, or demand further investigation. In terms of the relationship between prosody and formality in Japanese, it is quite possible that there are additional acoustic cues to the different levels of formality beyond the ones examined in Chapter 3. Analyses of properties of utterances

such as vowel quality, pause frequency, or amplitude (although it was not significant in the pilot in Chapter 2, it is still possible that it would be in speech from a corpus) might produce additional significant results which could be used to improve the accuracy of the probabilistic model in Chapter 5. It could also be useful to conduct additional studies similar to Chapter 3 cross-linguistically in order to re-evaluate current assumptions about the relationship between prosody and politeness (e.g. Brown & Levinson, 1987).

This thesis also leaves open some questions on the more specific nature of how prosody is used by speakers to express different levels of formality in conversation, as the rather large segments that were analyzed (full sentences rather than phrases, words, or moras) precluded a close investigation of exactly how prosody was changing at the segmental level. A more fine-grain investigation could help illuminate further whether the changes in prosody observed in this study are more closely connected to the phonological and morphological structure of Japanese. This study also examined speech data exclusively from speakers of Tokyo Japanese, but an investigation of how the changes in f_0 seen in this study interact with the pitch accent systems of different dialects could also help provide a better understanding of how the prosodic expression of formality is connected to the overall prosodic structure of Japanese.

As has been discussed in Section 6.2.2.2, this thesis also leaves open some questions on the specific cognitive mechanisms underlying the categorization task the subjects performed in

Chapter 4. The super-additive nature of the results also leaves open the more general question of whether it is possible for changes in any individual cue to have a significant effect on listeners' category judgments, or if such an effect will *only* be seen when cues are manipulated together. These questions could be addressed by a study which examined not only the effect of manipulation of these cues, but also the effect of the *degree* of manipulation. This approach, combined with more careful control of the unmanipulated base values of the phonetic parameters in question could provide further insight into the processes underlying the formal/informal category judgement task.

6.5 Conclusion

This experiments in this thesis investigating the relationship between prosody and formality have shown a number of novel and significant results. Foremost among these is compelling evidence that there is a consistent and significant relationship between the prosody of an utterance in Japanese and its formality, and also that the nature of this relationship – where f_0, articulation rate, and f_0 range all increase in informal speech – runs counter to previous results which had found a higher f_0 in polite speech in some contexts (Loveday, 1981, Ofuka et al., 2000). It was also shown in Chapter 4 that listeners are able to make use of their knowledge of this relationship (whether conscious or unconscious) in order to judge whether speech is formal or informal

without the presence of any lexical information. The speech perception study also produced a probabilistic theory of speech perception, which formed the basis for the structure of a successful predictive model of formality in Japanese. The methodologies used both for data collection and analysis were effective, showing significant results that were not seen consistently in previous similar studies of Japanese when less natural speech was analyzed.

While this thesis focused on one very specific theme – how different levels of formality can be expressed or understood in Japanese using prosody – the approaches taken throughout the thesis are not limited to investigating only that subject. The same approaches can be used to study formality cross-linguistically, and the probabilistic model in Chapter 5 could be used to model the relationship between any number of variables and any binary category. In and of itself, the work contained within the thesis has demonstrated that the relationship between formality in prosody in Japanese is both more consistent and more important than was previously known, and that an approach focused on making use of natural speech materials – and all the variation that entails – can produce significant results that otherwise might not be found.

Appendices

Appendix I – Model Coefficients

1.1 Chapter 2

Intensity Model (Full)

```
$pair
(Intercept)  FormalityF  ConditionT    GenderM      FormalityF : ConditionT
1  67.8441   0.1187     -0.1821        6.5183       0.0555
2  68.5916   0.0606     -0.1821        6.5183       0.0555
3  68.1589   0.0943     -0.1821        6.5183       0.0555
4  73.4242  -0.3150     -0.1821        6.5183       0.0555
5  71.8385  -0.1918     -0.1821        6.5183       0.0555
6  73.5186  -0.3224     -0.1821        6.5183       0.0555
7  69.2099   0.0125     -0.1821        6.5183       0.0555
8  68.4872   0.0687     -0.1821        6.5183       0.0555
9  71.9388  -0.1995     -0.1821        6.5183       0.0555
10 68.4796   0.0693     -0.1821        6.5183       0.0555
11 69.4906  -0.0092     -0.1821        6.5183       0.0555
12 72.1928  -0.2193     -0.1821        6.5183       0.0555
13 67.7501   0.1260     -0.1821        6.5183       0.0555
14 68.1800   0.0926     -0.1821        6.5183       0.0555

$speaker
(Intercept)  FormalityF  ConditionT    GenderM      FormalityF : ConditionT
1  73.05529  0.1355     -0.1821        6.5183       0.0555
2  70.08355 -0.0353     -0.1821        6.5183       0.0555
3  69.27551 -0.0818     -0.1821        6.5183       0.0555
4  67.50332 -0.1837     -0.1821        6.5183       0.0555
5  69.76287 -0.0538     -0.1821        6.5183       0.0555
```

Intensity Model (Treatment)

```
$speaker
(Intercept)  FormalityF   GenderM    FormalityF : GenderM
1  72.52840  0.1502       6.6272     0.2398
2  69.83987 -0.0748       6.6272     0.2398
3  69.43761 -0.1085       6.6272     0.2398
4  66.89235 -0.3217       6.6272     0.2398
5  69.85365 -0.0737       6.6272     0.2398

$pair
(Intercept)  FormalityF   GenderM    FormalityF : GenderM
3  67.99808 -0.0588       6.6272     0.2398
7  68.99457 -0.0745       6.6272     0.2398
9  71.65766 -0.1162       6.6272     0.2398
12 71.89029 -0.1199       6.6272     0.2398
14 68.01130 -0.0590       6.6272     0.2398
```

F_0 Model (Interaction)

```
$pair
(Intercept)  FormalityF   ConditionT    GenderM      FormalityF : ConditionT
1  251.8798  3.9349       6.3388       -111.9817    -15.5599
2  265.1545  2.7798       6.3388       -111.9817    -15.5599
3  262.5752  2.5517       6.3388       -111.9817    -15.5599
4  266.6398  1.3002       6.3388       -111.9817    -15.5599
5  286.2443  0.3936       6.3388       -111.9817    -15.5599
6  273.1537  0.1995       6.3388       -111.9817    -15.5599
7  277.6596 -1.1178       6.3388       -111.9817    -15.5599
8  263.4458 -1.5994       6.3388       -111.9817    -15.5599
9  268.9549  2.4224       6.3388       -111.9817    -15.5599
10 275.9011 -1.0943       6.3388       -111.9817    -15.5599
11 271.9571  1.0259       6.3388       -111.9817    -15.5599
12 271.2015 -0.9942       6.3388       -111.9817    -15.5599
13 286.7889 -0.4046       6.3388       -111.9817    -15.5599
14 275.8116  0.7688       6.3388       -111.9817    -15.5599

$speaker
(Intercept)  FormalityF   ConditionT    GenderM      FormalityF : ConditionT
1  289.0188  1.0661       6.3388       -111.9817    -15.5599
2  289.6718  1.0786       6.3388       -111.9817    -15.5599
3  263.3503  0.5753       6.3388       -111.9817    -15.5599
4  252.0725  0.3596       6.3388       -111.9817    -15.5599
5  262.0893  0.5511       6.3388       -111.9817    -15.5599
```

F_0 Model (Treatment)

```
$speaker
(Intercept)  FormalityF   GenderM        FormalityF : GenderM
2  297.4370 -15.1537     -106.6687      -0.8557
3  291.6523 -14.9992     -106.6687      -0.8557
4  267.3266 -14.3494     -106.6687      -0.8557
5  253.4713 -13.9793     -106.6687      -0.8557
6  267.3836 -14.3509     -106.6687      -0.8557

$pair
(Intercept)  FormalityF   GenderM        FormalityF : GenderM
3  266.5396 -10.78249    -106.6687      -0.8557
7  282.0661 -17.54782    -106.6687      -0.8557
9  272.9897 -13.11878    -106.6687      -0.8557
12 275.5404 -15.18582    -106.6687      -0.8557
14 280.1351 -16.19789    -106.6687      -0.8557
```

Duration Model (Interaction)

```
$pair
(Intercept)  FormalityF   ConditionT    GenderM      FormalityF : ConditionT
1  492.6784 -12.8420     -17.0140      -99.7699      14.3407
2  436.0667  21.0545     -17.0140      -99.7699      14.3407
```

```
3  485.6716  29.2689     -17.0140      -99.7699      14.3407
4  643.0778  41.6401     -17.0140      -99.7699      14.3407
5  727.6215   5.2818     -17.0140      -99.7699      14.3407
6  780.6774   5.7461     -17.0140      -99.7699      14.3407
7  461.2194  13.1781     -17.0140      -99.7699      14.3407
8  502.4286  35.3045     -17.0140      -99.7699      14.3407
9  664.6240  25.7267     -17.0140      -99.7699      14.3407
10  622.6666  23.5516    -17.0140      -99.7699      14.3407
11  572.3254  18.2492    -17.0140      -99.7699      14.3407
12  767.6183  14.2872    -17.0140      -99.7699      14.3407
13  413.6754  -8.6321    -17.0140      -99.7699      14.3407
14  504.8765 -10.5978    -17.0140      -99.7699      14.3407

$speaker
(Intercept)  FormalityF   ConditionT    GenderM      FormalityF : ConditionT
1  577.1416  14.0256      -17.0140      -99.7699      14.3407
2  606.3995  26.9628      -17.0140      -99.7699      14.3407
3  536.1840  -2.3134      -17.0140      -99.7699      14.3407
4  582.8789  17.4977      -17.0140      -99.7699      14.3407
5  581.4058  15.6905      -17.0140      -99.7699      14.3407
```

Duration Model (Treatment)

```
$speaker
(Intercept)  FormalityF   GenderM        FormalityF : GenderM
1  552.1283  18.5783     -85.8833        23.0541
2  588.1527  30.6992     -85.8833        23.0541
3  510.0474   4.4197     -85.8833        23.0541
4  556.3384  19.9949     -85.8833        23.0541
5  564.4999  22.7409     -85.8833        23.0541

$pair
(Intercept)  FormalityF   GenderM        FormalityF : GenderM
3  463.4921  33.4934     -85.8833        23.0541
7  438.5569  17.9107     -85.8833        23.0541
9  642.2799  29.9144     -85.8833        23.0541
12 745.1301  19.5243     -85.8833        23.0541
14 481.7076  -4.4097     -85.8833        23.0541
```

Duration Model (Control)

```
$pair
(Intercept)  FormalityF   GenderM        FormalityF : GenderM
1  493.3944 -10.9598     -100.4681        -6.7355
2  436.2434  24.0350     -100.4681        -6.7355
4  642.8160  45.2462     -100.4681        -6.7355
5  728.0370   7.6175     -100.4681        -6.7355
6  781.0830   8.0697     -100.4681        -6.7355
8  502.3743  38.7013     -100.4681        -6.7355
10 622.7948  26.5168     -100.4681        -6.7355
11 572.5406  21.0724     -100.4681        -6.7355
13 414.3277  -6.5772     -100.4681        -6.7355

$speaker
(Intercept)  FormalityF   GenderM        FormalityF : GenderM
1  573.6394  15.06161    -100.4681        -6.7355
```

```
2  606.6183  31.61597    -100.4681       -6.7355
3  542.9829   1.14554    -100.4681       -6.7355
4  580.4964  19.09889    -100.4681       -6.7355
5  581.6026  18.47923    -100.4681       -6.7355
```

1.2 Chapter 3

Log_{10} F_0 Model

```
$speaker
   (Intercept)   FormalityF    GenderM        FormalityF : GenderM
1   2.4087      -0.03405      -0.22127        -0.00243
2   2.3963      -0.03456      -0.22127        -0.00243
3   2.2877      -0.03910      -0.22127        -0.00243
4   2.3282      -0.03741      -0.22127        -0.00243
5   2.2389      -0.04115      -0.22127        -0.00243
6   2.2592      -0.04030      -0.22127        -0.00243
7   2.2992      -0.03862      -0.22127        -0.00243
8   2.3615      -0.03602      -0.22127        -0.00243
9   2.2087      -0.04241      -0.22127        -0.00243
10  2.2834      -0.03928      -0.22127        -0.00243
```

Articulation Rate Model

```
$speaker
   (Intercept)   FormalityF    GenderM         FormalityF : GenderM
1   6.5365      -0.8169       -0.17266        -0.20374
2   7.7059      -0.9220       -0.17266        -0.20374
3   7.3498      -1.0323       -0.17266        -0.20374
4   7.1377      -0.9735       -0.17266        -0.20374
5   8.1047      -1.1567       -0.17266        -0.20374
6   7.8955      -0.9965       -0.17266        -0.20374
7   8.4839      -1.3760       -0.17266        -0.20374
8   8.8751      -1.2369       -0.17266        -0.20374
9   8.8607      -1.2939       -0.17266        -0.20374
10  8.5341      -1.4258       -0.17266        -0.20374
```

F_0 Range Model

```
$speaker
   (Intercept)   FormalityF    GenderM          FormalityF : GenderM
1   173.5597    -37.5881      -36.09306         7.06184
2   179.2257    -43.5370      -36.09306         7.06184
3   162.8431    -38.3386      -36.09306         7.06184
4   157.3246    -39.0440      -36.09306         7.06184
5   131.3182    -39.7578      -36.09306         7.06184
```

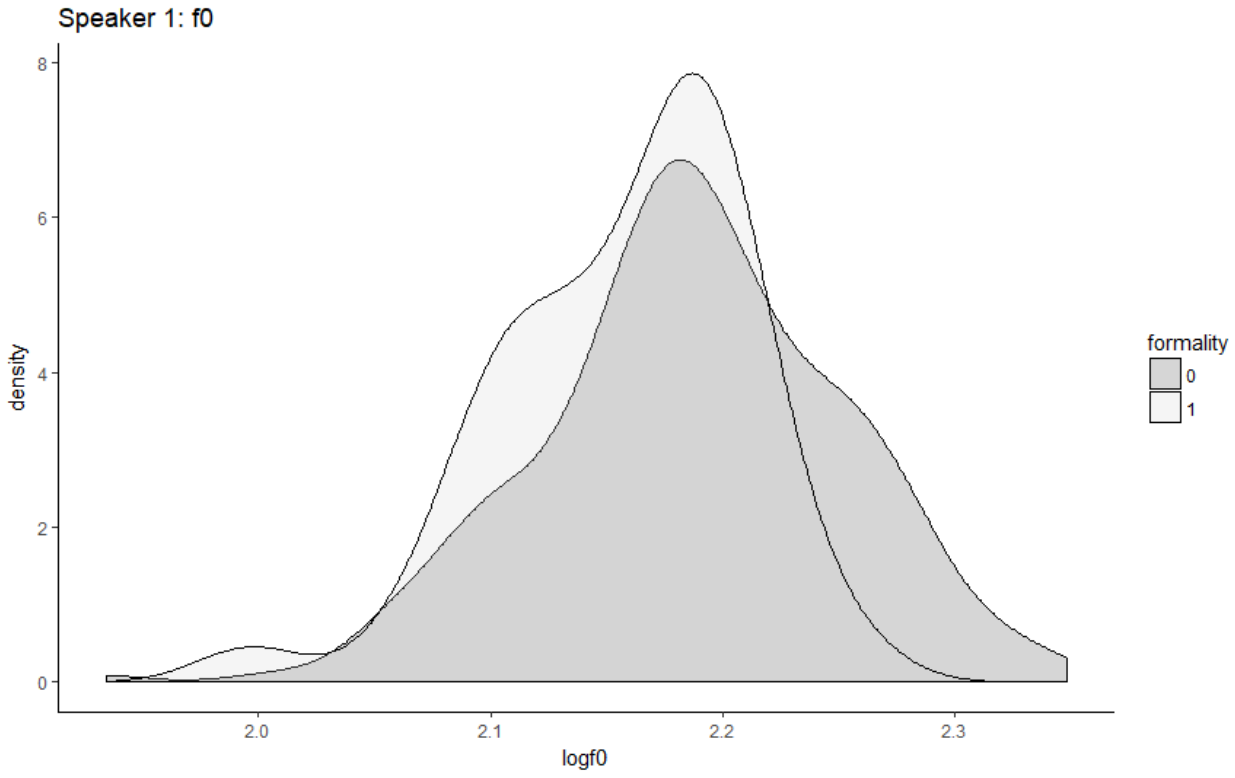
```
6   156.7384    -39.3390      -36.09306         7.06184
7   124.6116    -38.8924      -36.09306         7.06184
8   139.9780    -42.3202      -36.09306         7.06184
9   115.2764    -36.4535      -36.09306         7.06184
10  142.1159    -40.5977      -36.09306         7.06184
```

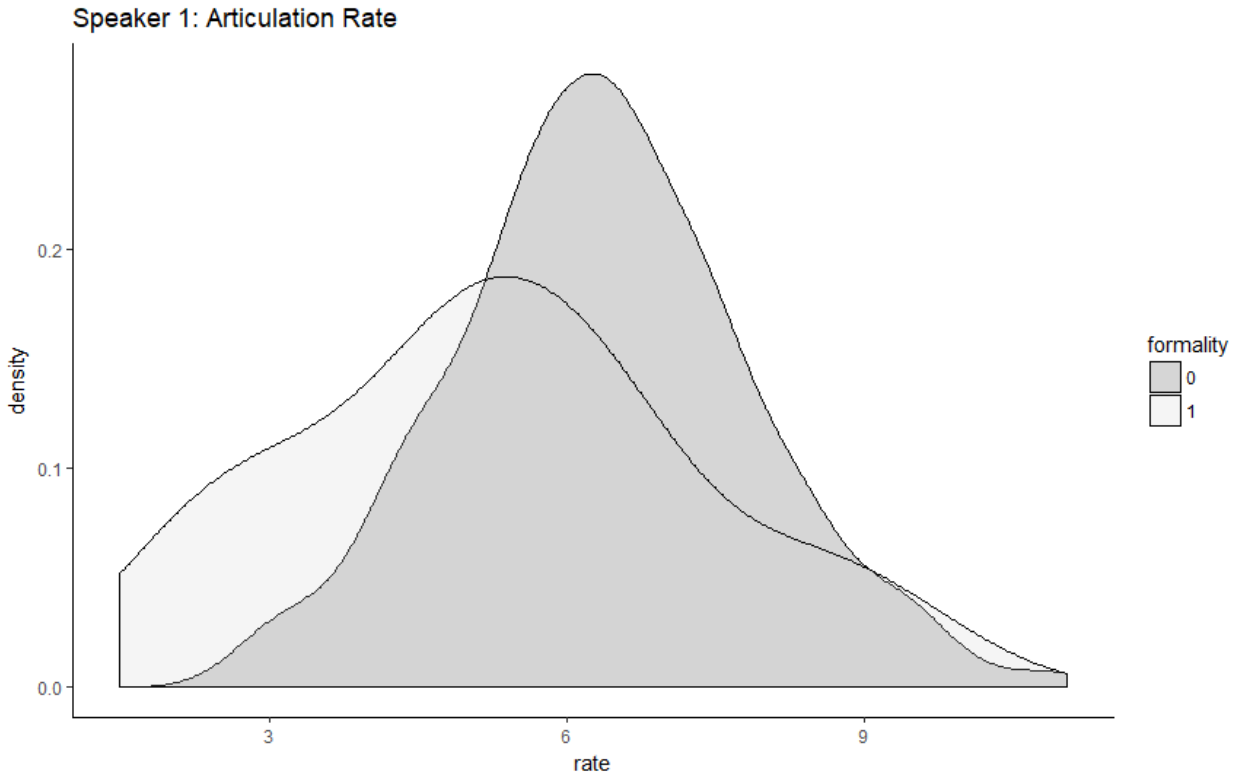
Function Coefficients Model

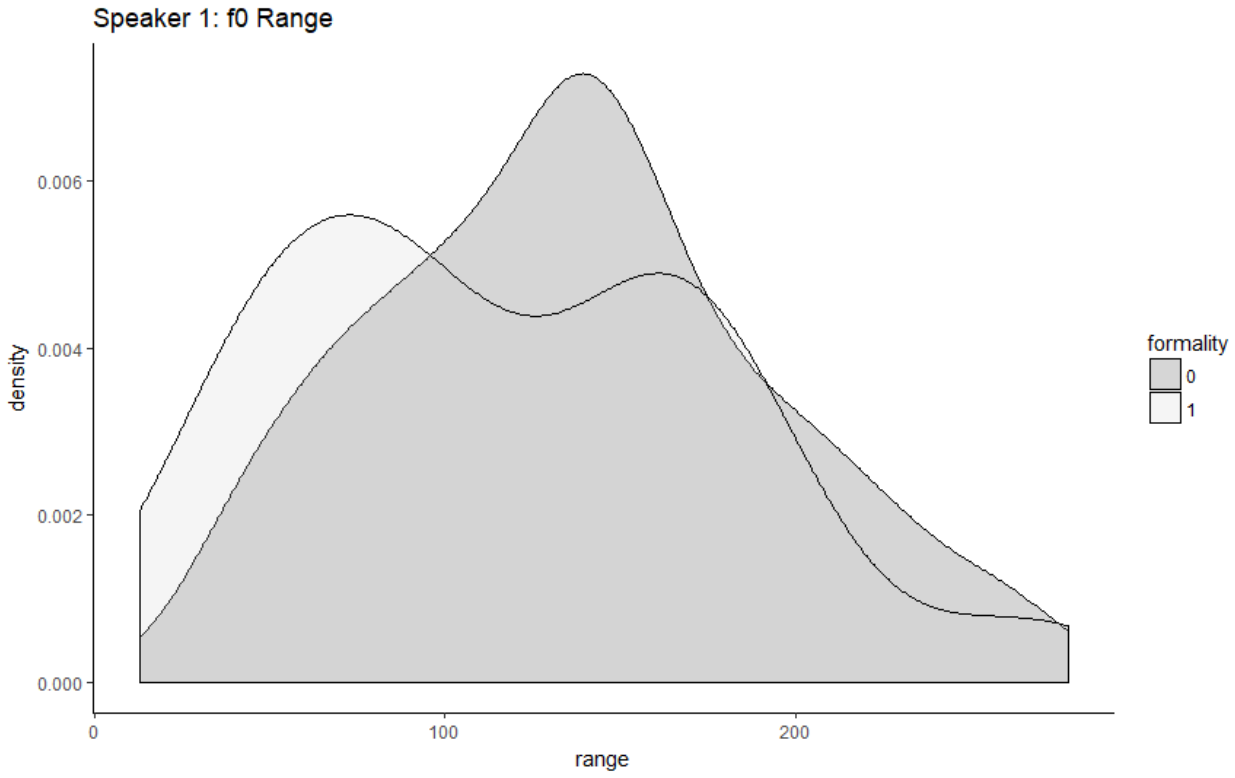
```
$speaker
   (Intercept)   c1         c2         c3         c4
1    -1.869264  -0.02738    0.31408   -0.42860  -0.03390
2    -1.197273  -0.14404    1.28595   -0.18019  -0.65194
3    -1.624169   0.07154    0.72340   -0.35764  -0.17740
4    -1.654406   0.04009    0.32927   -0.26751  -0.08514
5    -1.695660   0.07670    0.35482   -0.30797  -0.04896
6    -1.651766   0.09514    0.40032   -0.28732  -0.06708
7    -1.651495  -0.03490    0.33078   -0.26358  -0.13762
8    -1.418693  -0.06351    0.88712   -0.24072  -0.39972
9    -1.573150   0.19648    0.01218   -0.11784   0.06647
10   -1.423002  -0.03767    0.70973   -0.19403  -0.33512
```

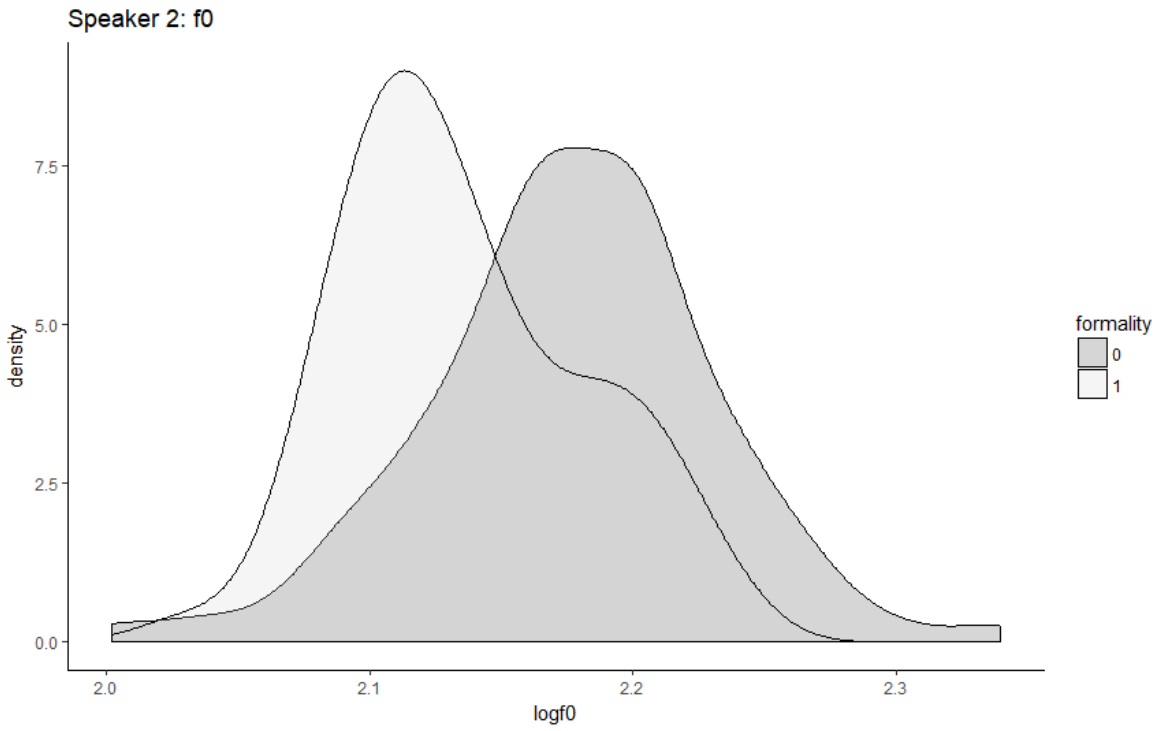
Appendix II – Individual Speaker Density Plots (Chapter 3)

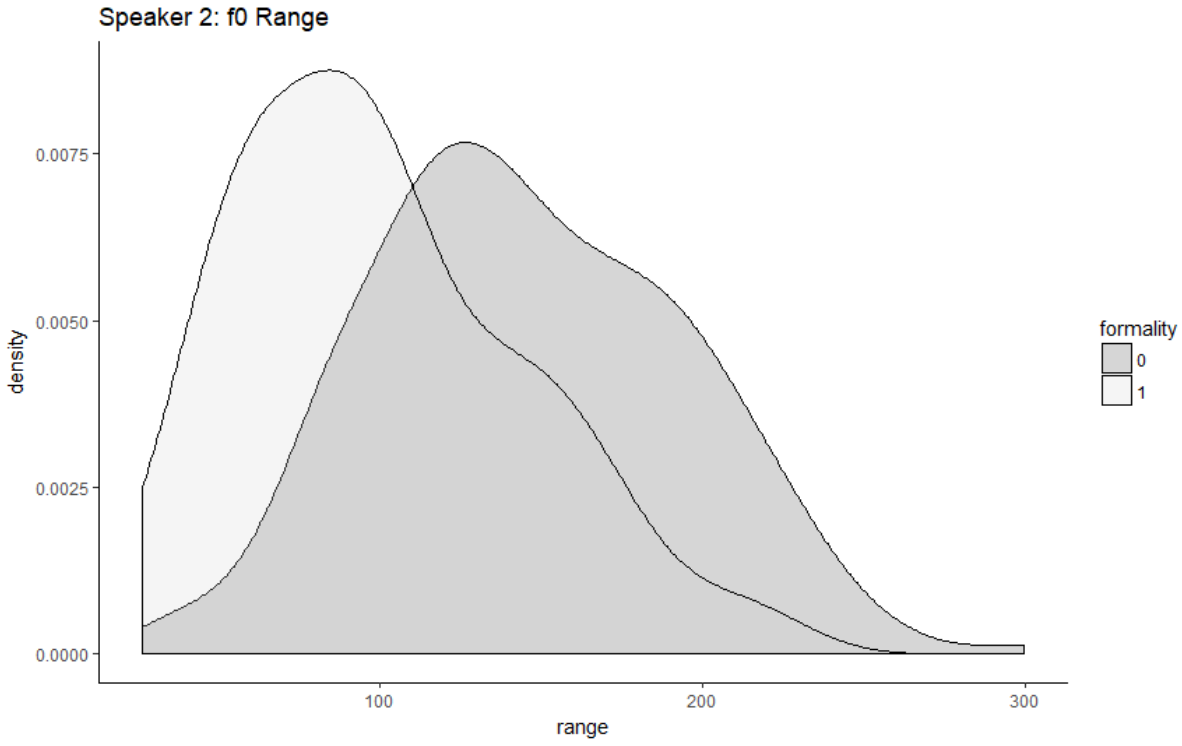
Density plots of each variable for each speaker broken down by formality. Formality 0 is *informal,* formality 1 is *formal.* Plots are grouped by speaker.

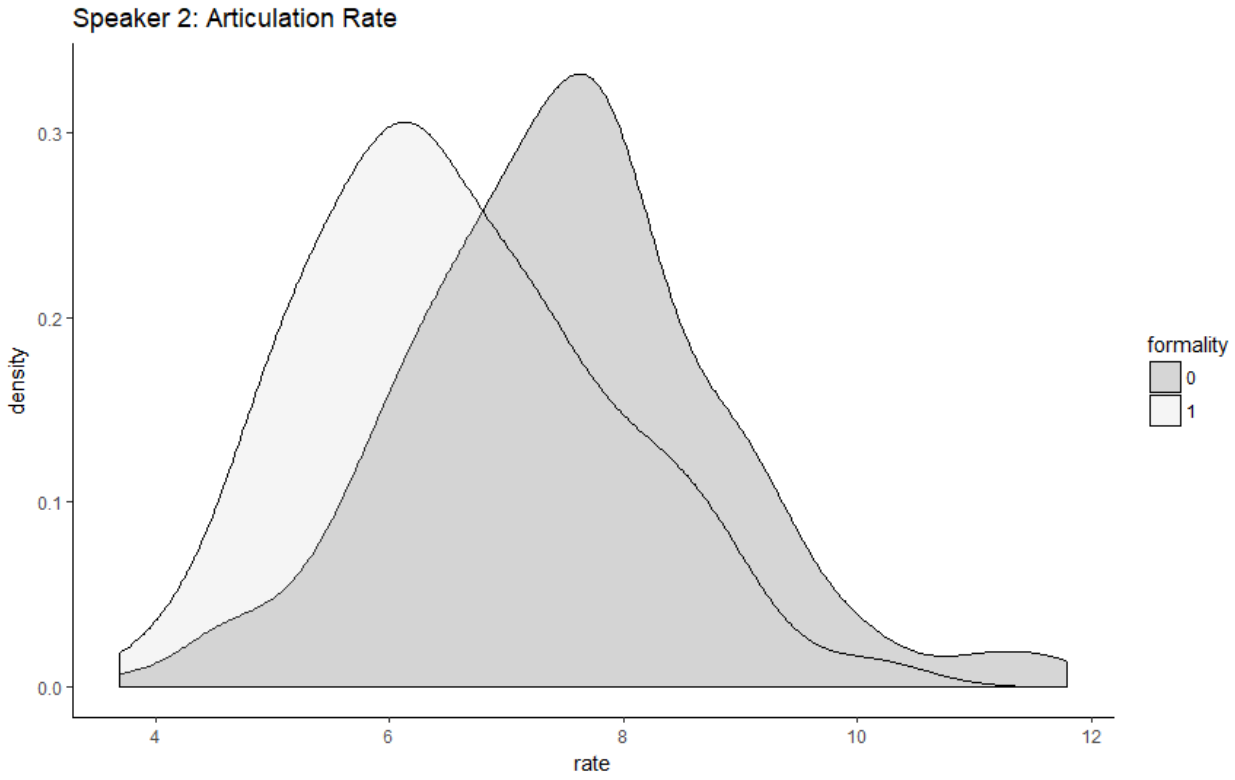


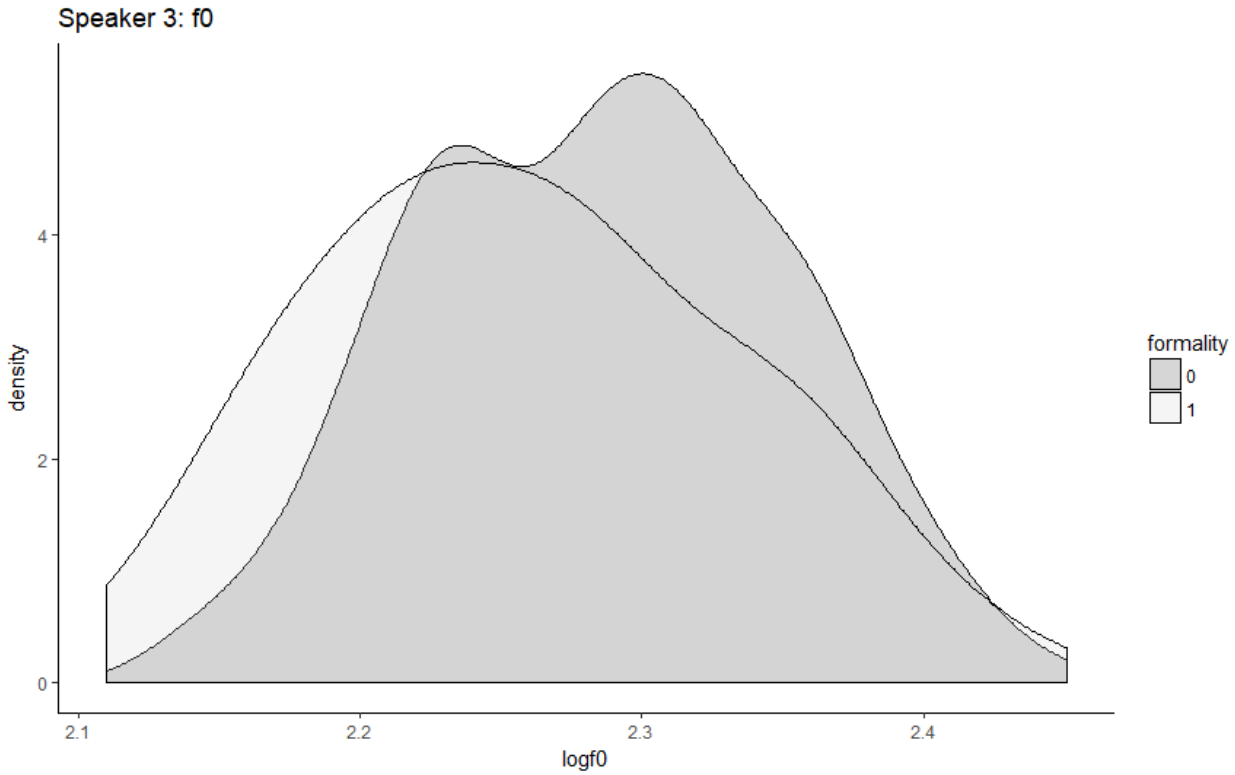

Speaker 1: Articulation Rate
density
0.2
0.1
0.0
3
6
9
rate
formality
0
1

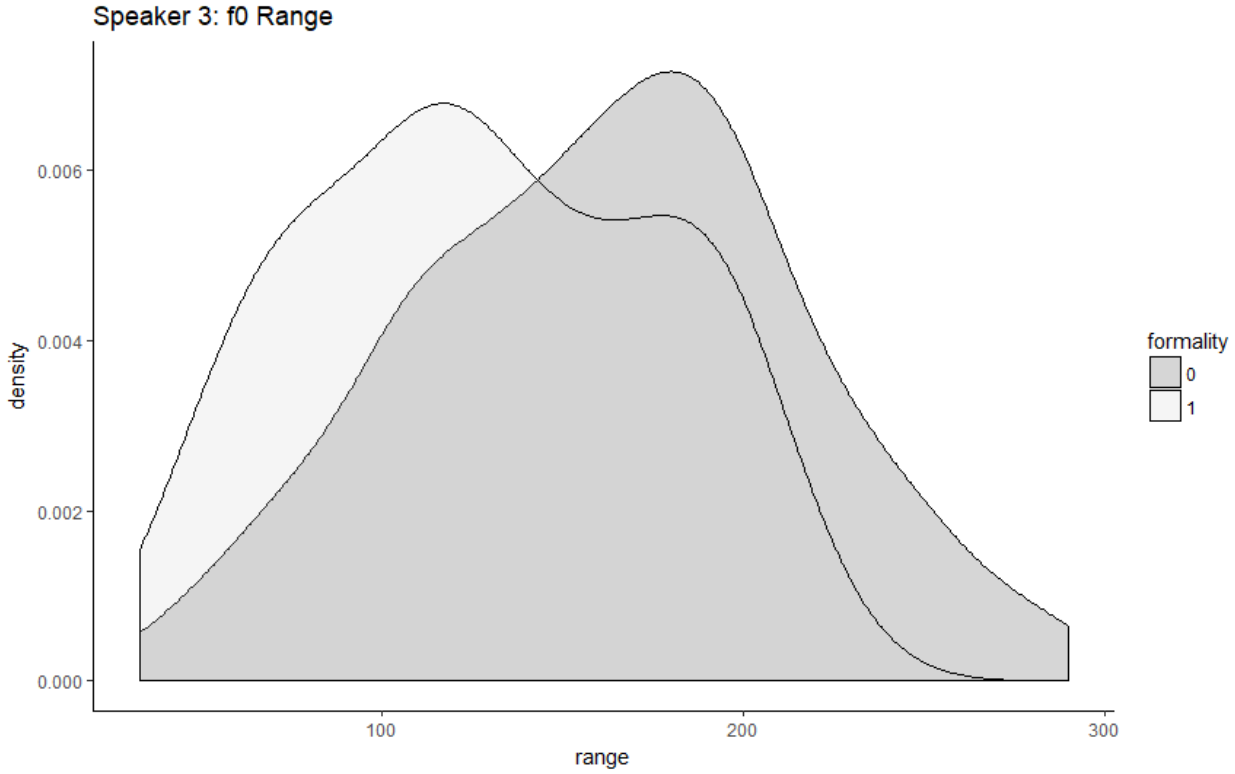

Speaker 1: f0 Range
density
0.006
0.004
0.002
0.000
0
100
200
range
formality
0
1

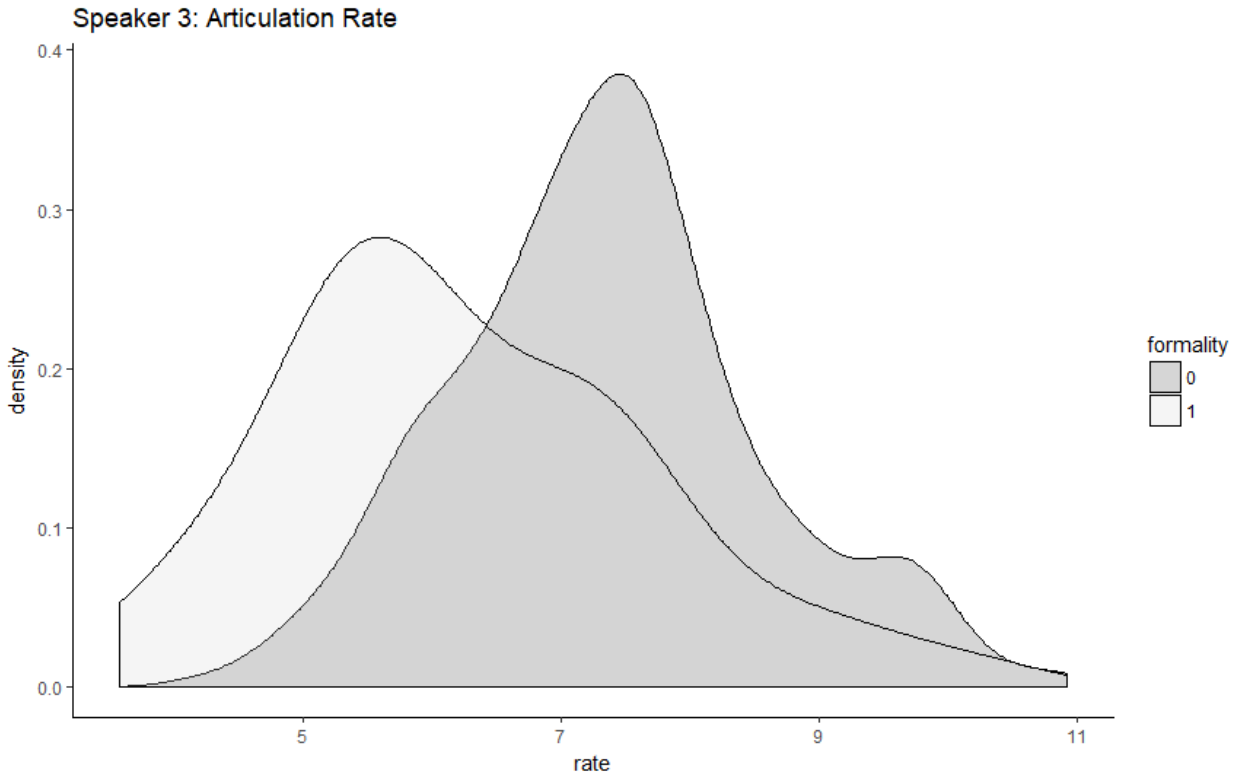

Speaker 2: f0
density
7.5
5.0
2.5
0.0
2.0
2.1
2.2
2.3
logf0
formality
0
1

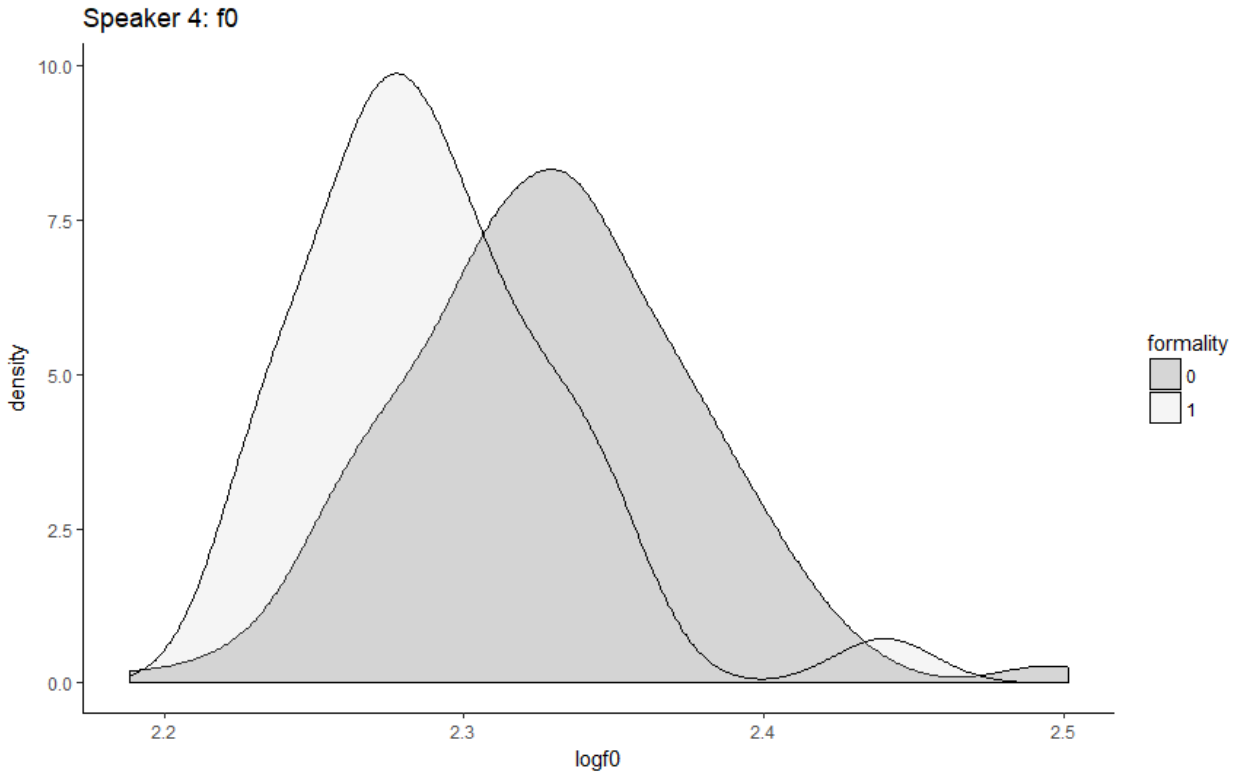

Speaker 2: f0 Range
density
0.0075
0.0050
0.0025
0.0000
100
200
300
range
formality
0
1

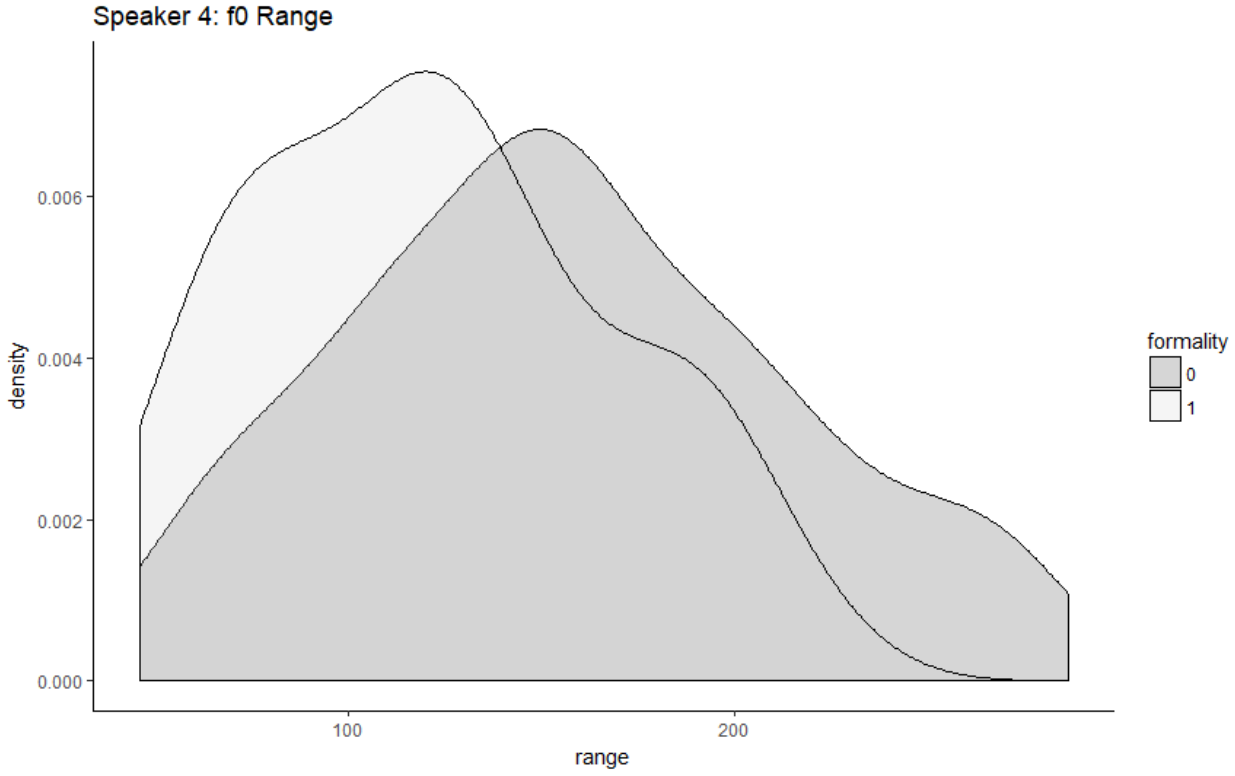

Speaker 2: Articulation Rate
density
0.0
0.1
0.2
0.3
4
6
8
10
12
rate
formality
0
1

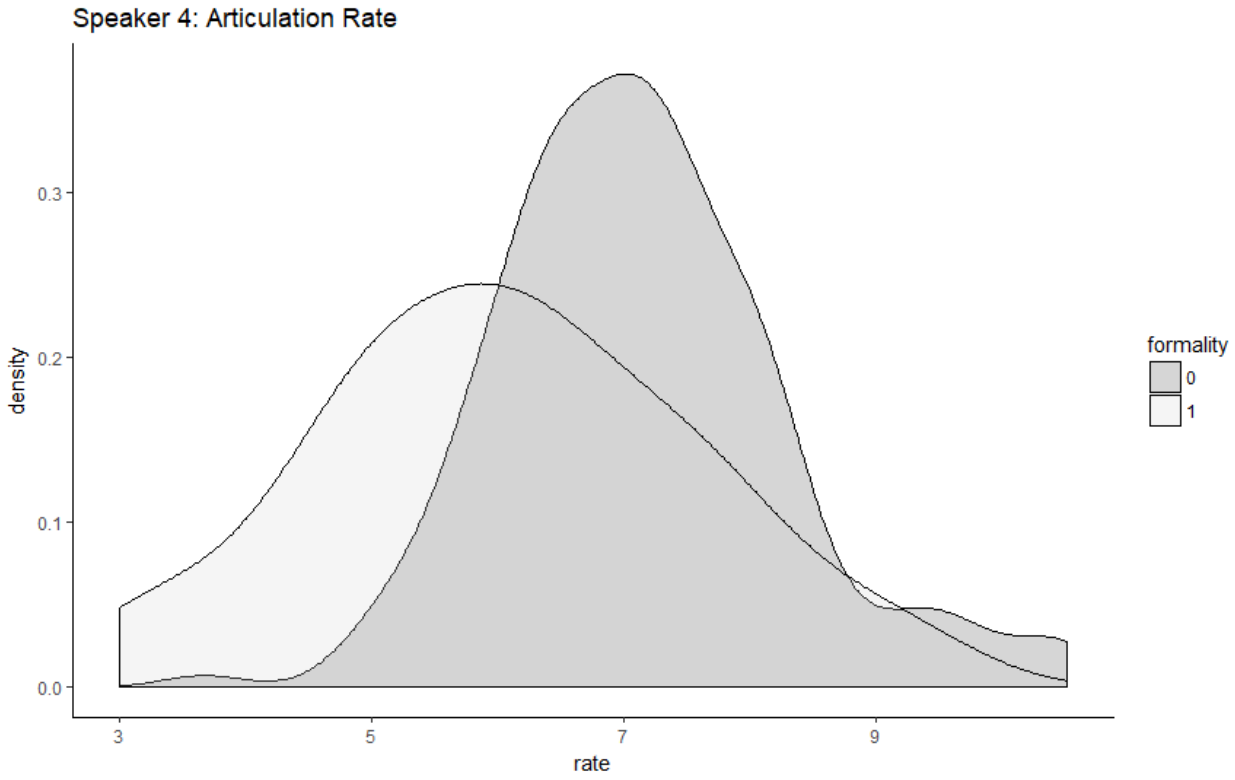

Speaker 3: f0
density
4
2
0
2.1
2.2
2.3
2.4
logf0
formality
0
1

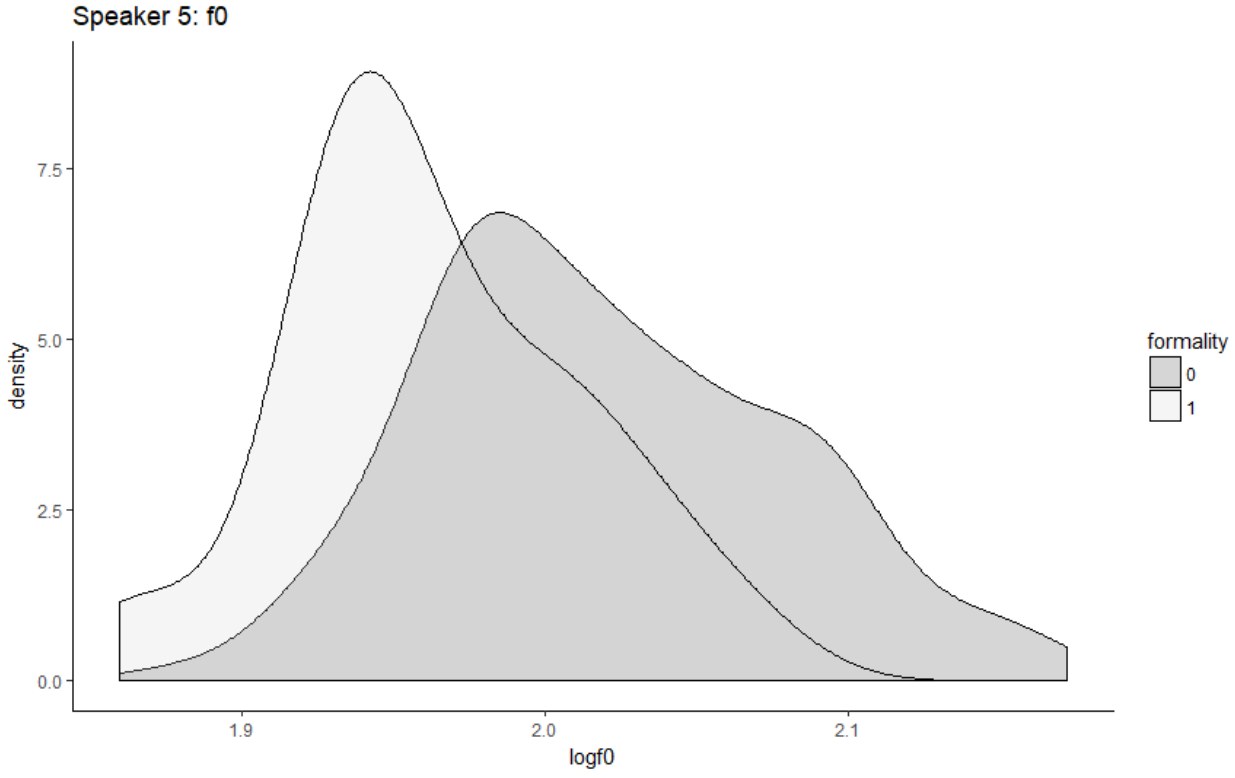

Speaker 3: f0 Range
density
0.006
0.004
0.002
0.000
100
200
300
range
formality
0
1

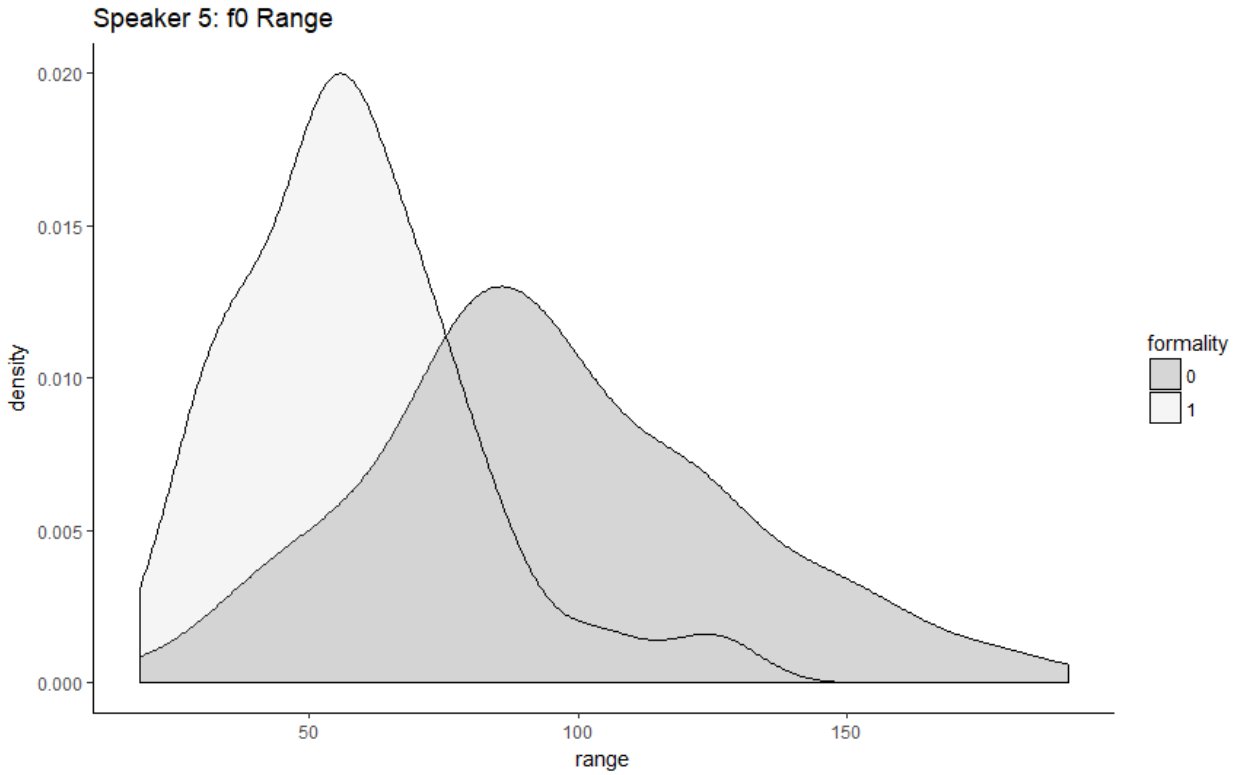

Speaker 3: Articulation Rate
density
0.4
0.3
0.2
0.1
0.0
5
7
9
11
rate
formality
0
1

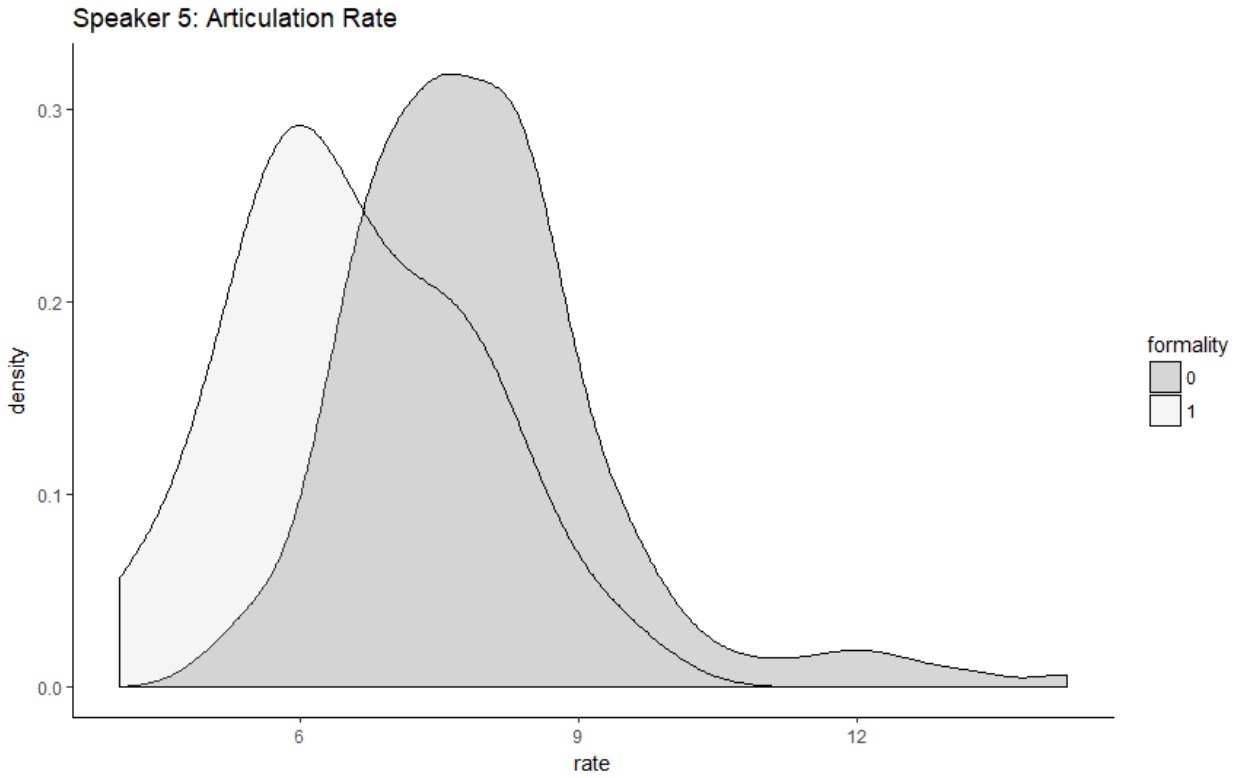

Speaker 4: f0
density
10.0
7.5
5.0
2.5
0.0
2.2
2.3
2.4
2.5
logf0
formality
0
1

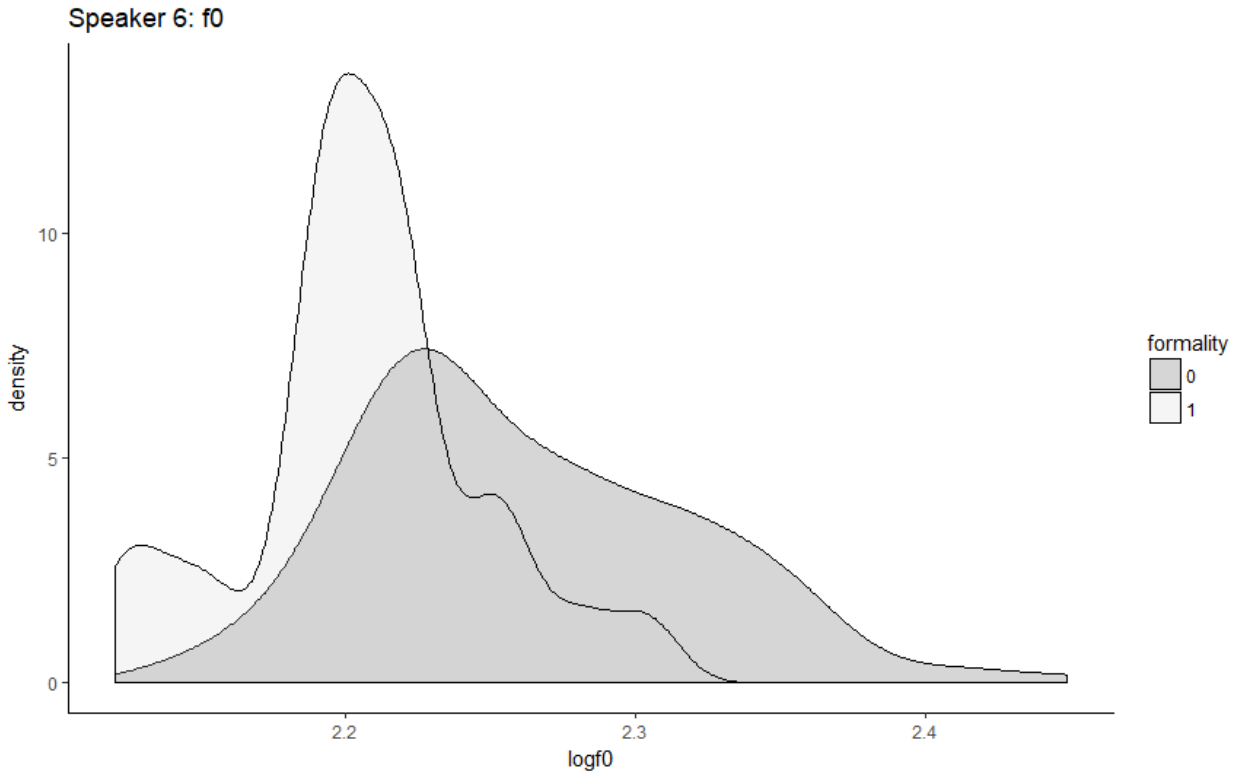

Speaker 4: f0 Range
density
0.006
0.004
0.002
0.000
100
200
range
formality
0
1

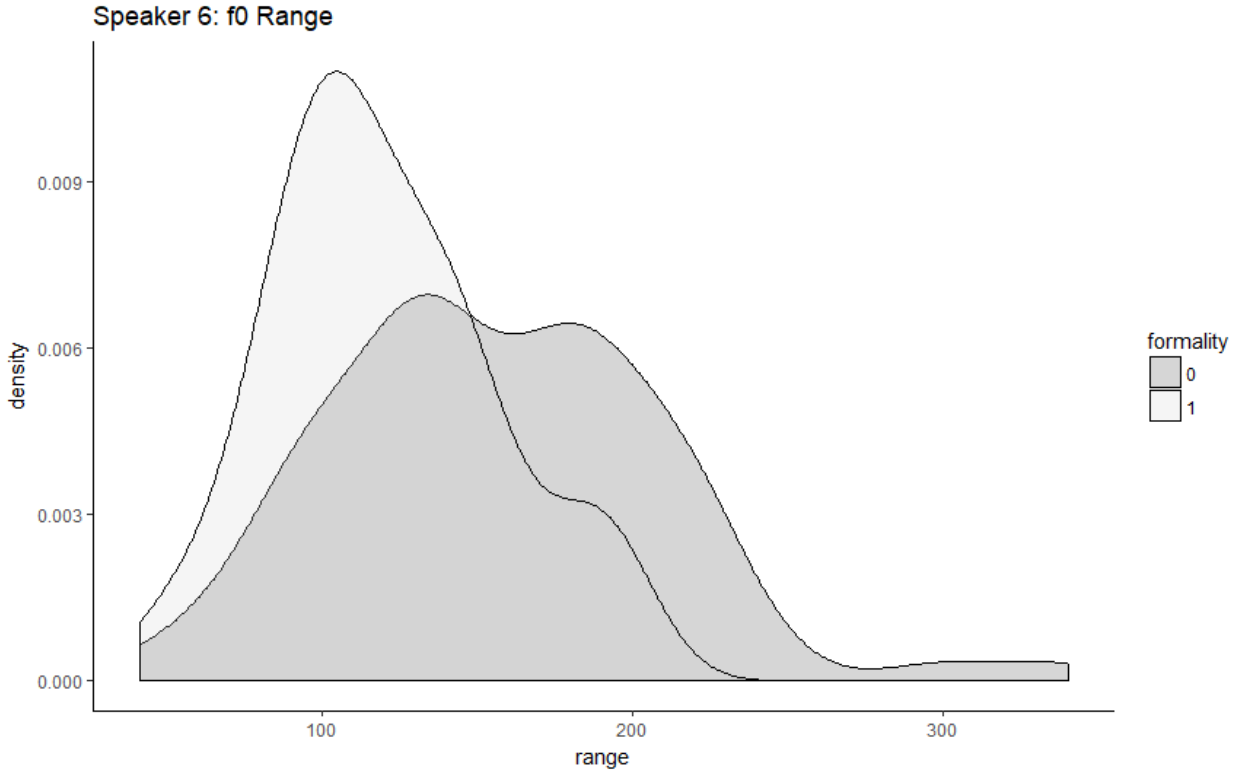

Speaker 4: Articulation Rate
density
0.0
0.1
0.2
0.3
3
5
7
9
rate
formality
0
1

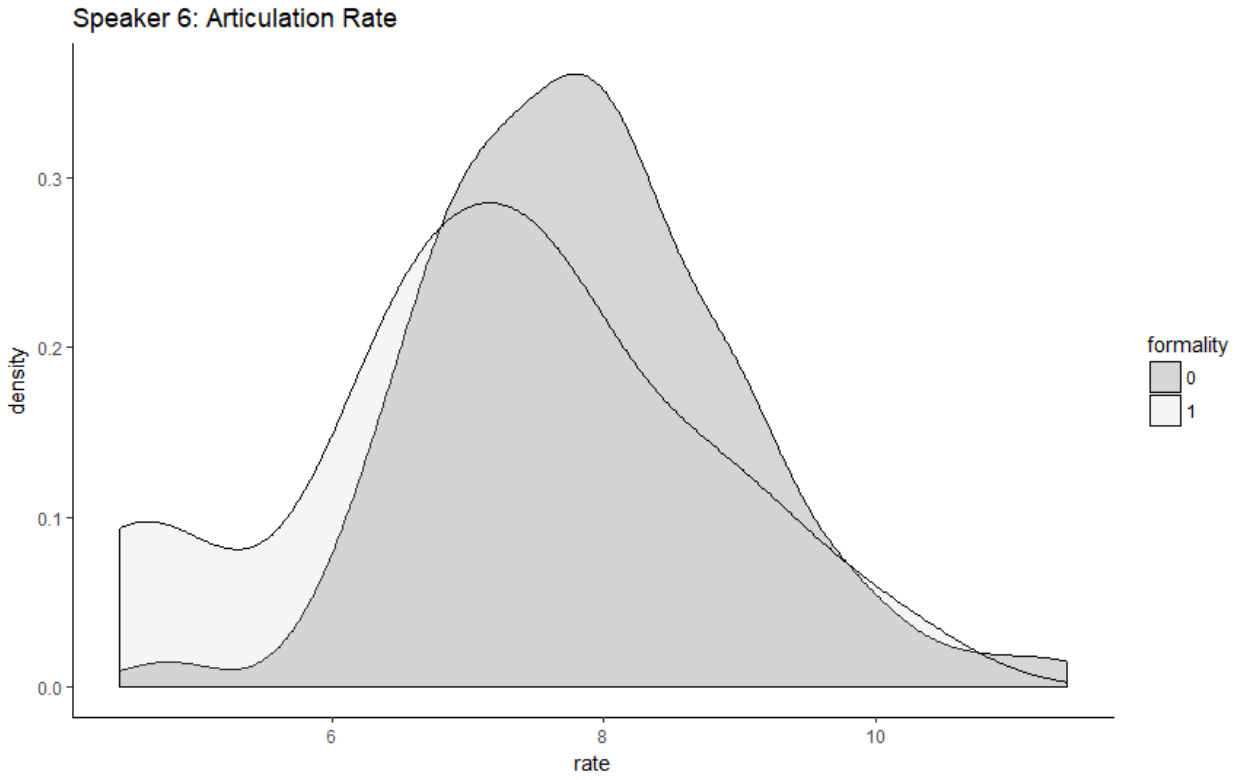

Speaker 5: f0
density
0.0
2.5
5.0
7.5
1.9
2.0
2.1
logf0
formality
0
1

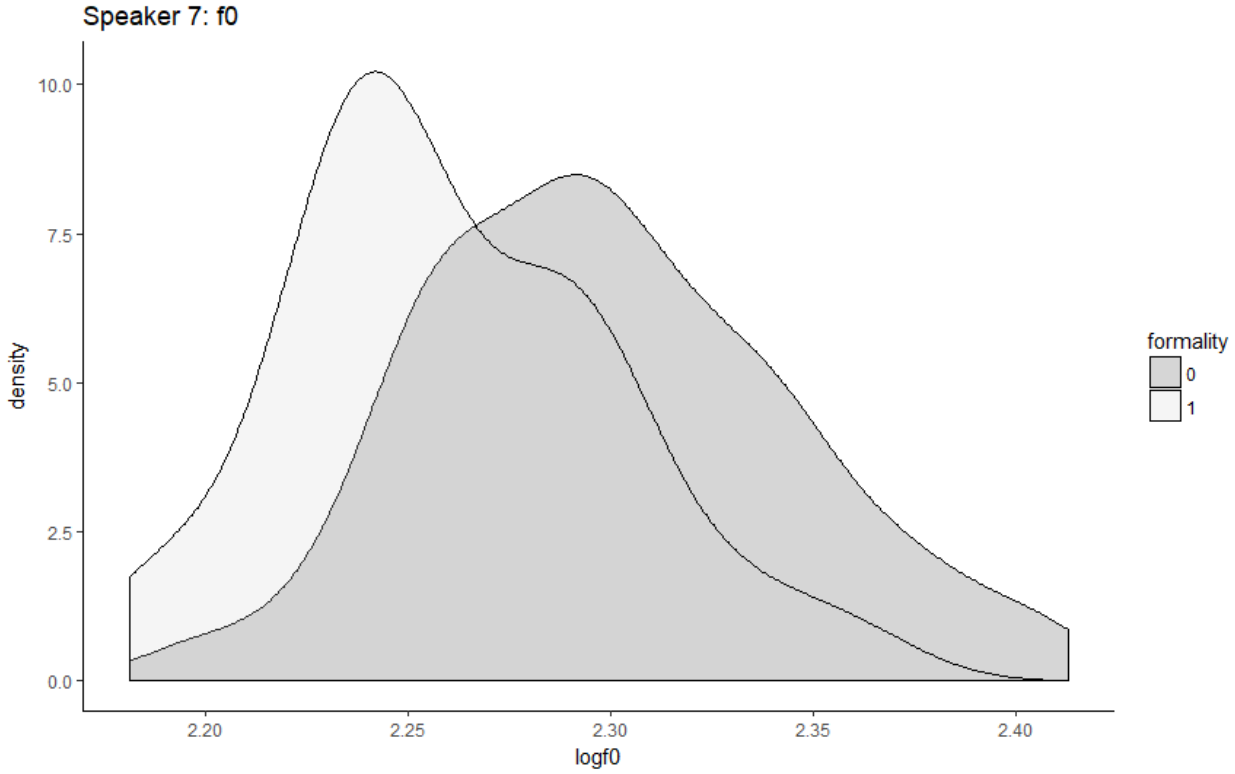

Speaker 5: f0 Range
density
0.020
0.015
0.010
0.005
0.000
50
100
150
range
formality
0
1

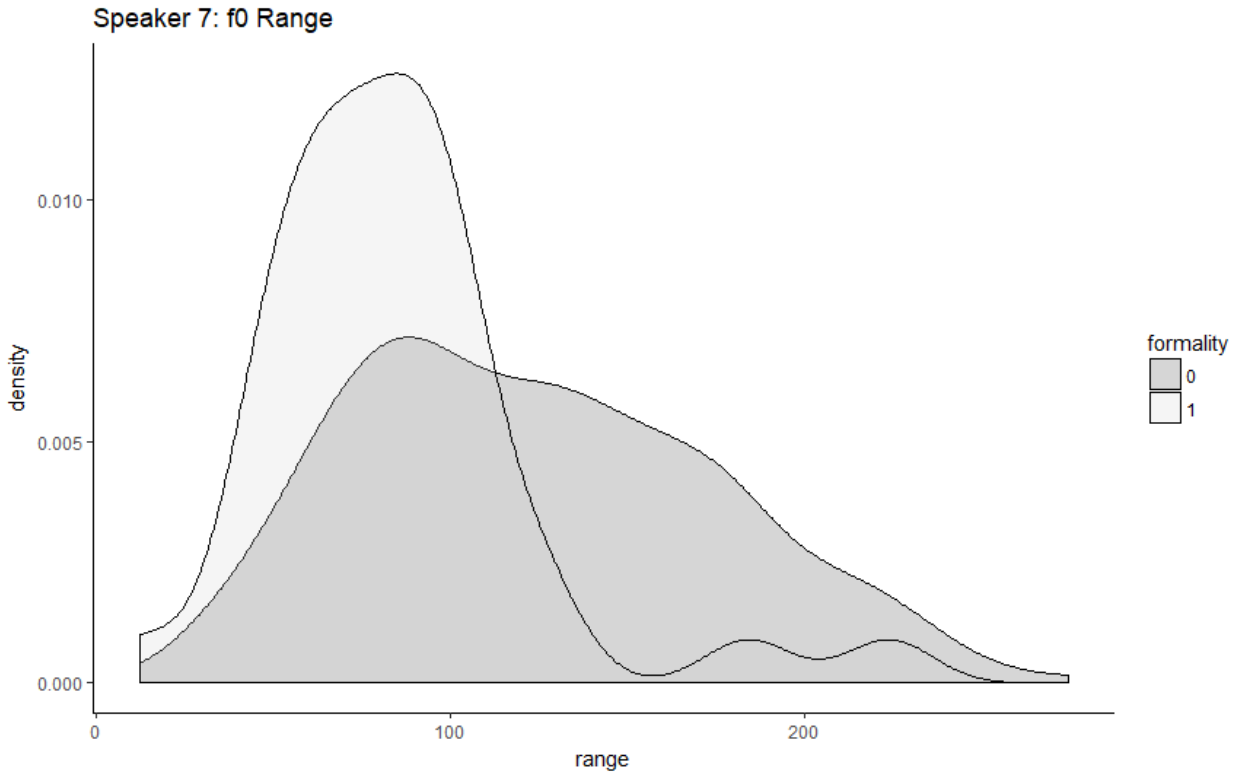

Speaker 5: Articulation Rate
density
0.3
0.2
0.1
0.0
6
9
12
rate
formality
0
1

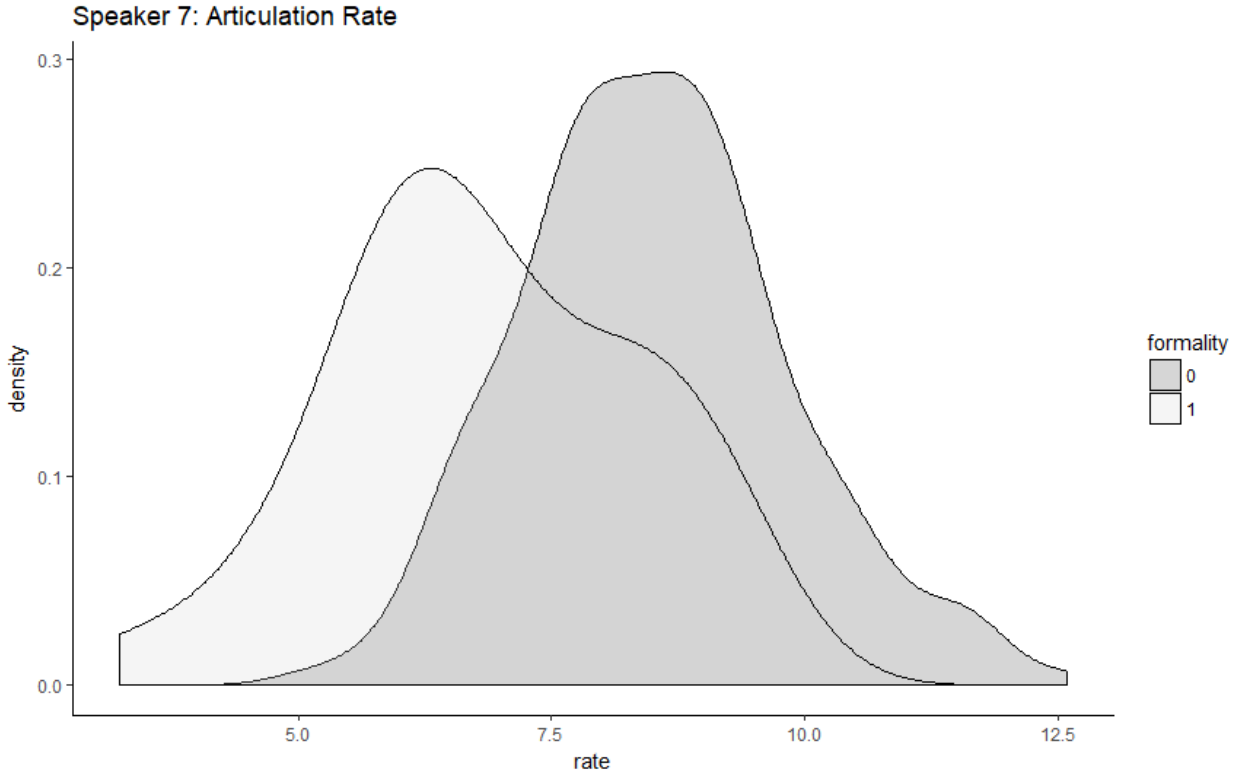

Speaker 6: f0
density
10
5
0
2.2
2.3
2.4
logf0
formality
0
1

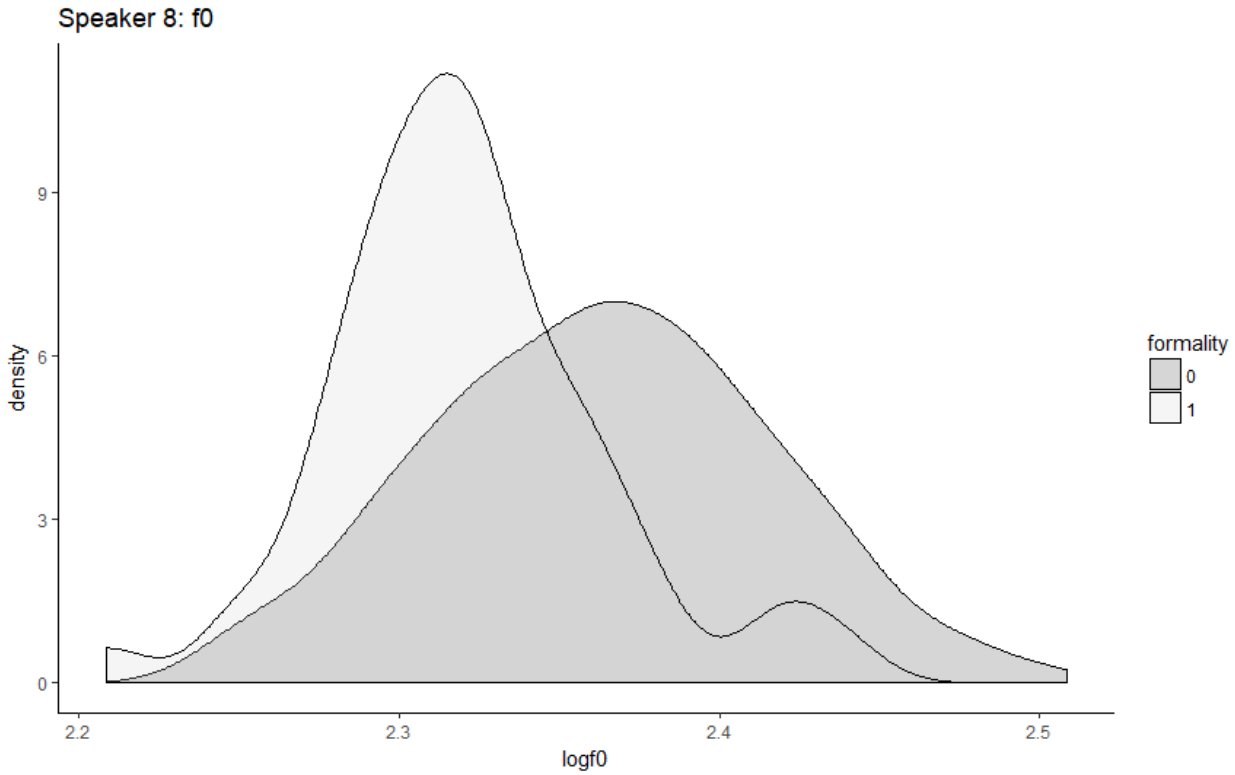

Speaker 6: f0 Range
density
0.009
0.006
0.003
0.000
100
200
300
range
formality
0
1

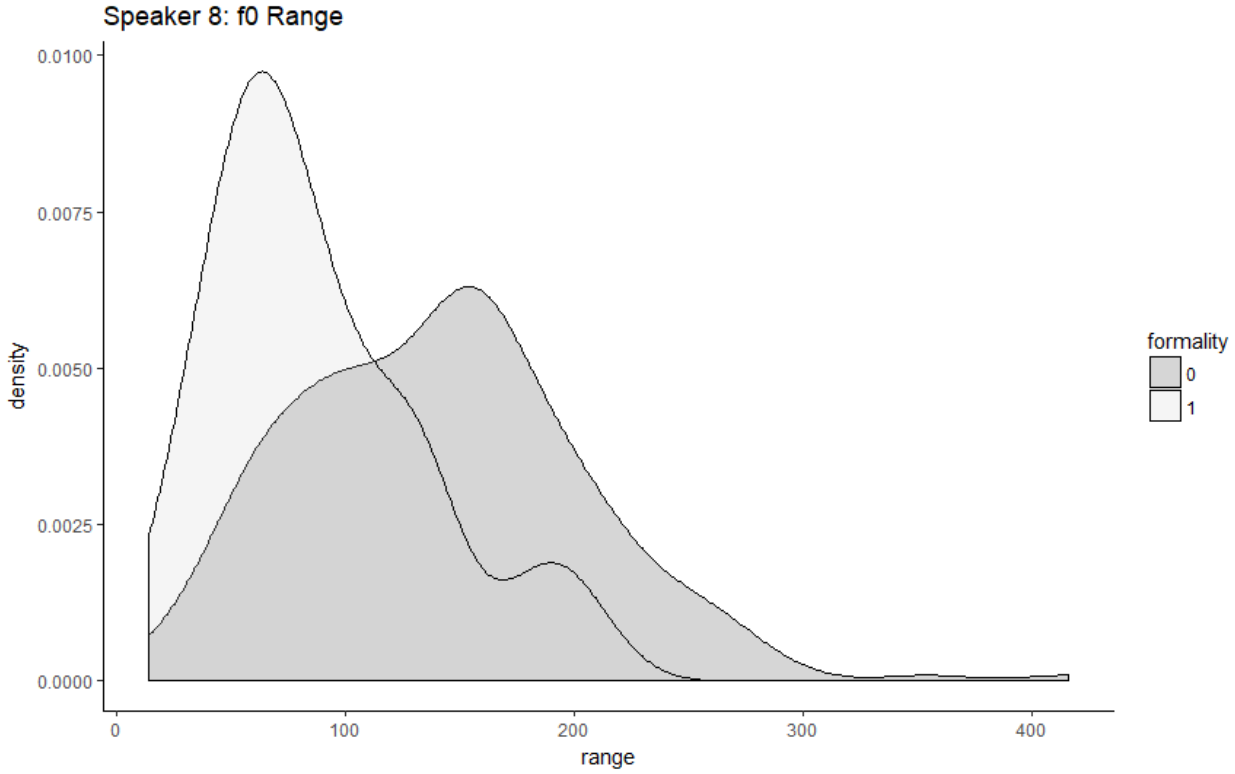

Speaker 6: Articulation Rate
density
0.0
0.1
0.2
0.3
6
8
10
rate
formality
0
1

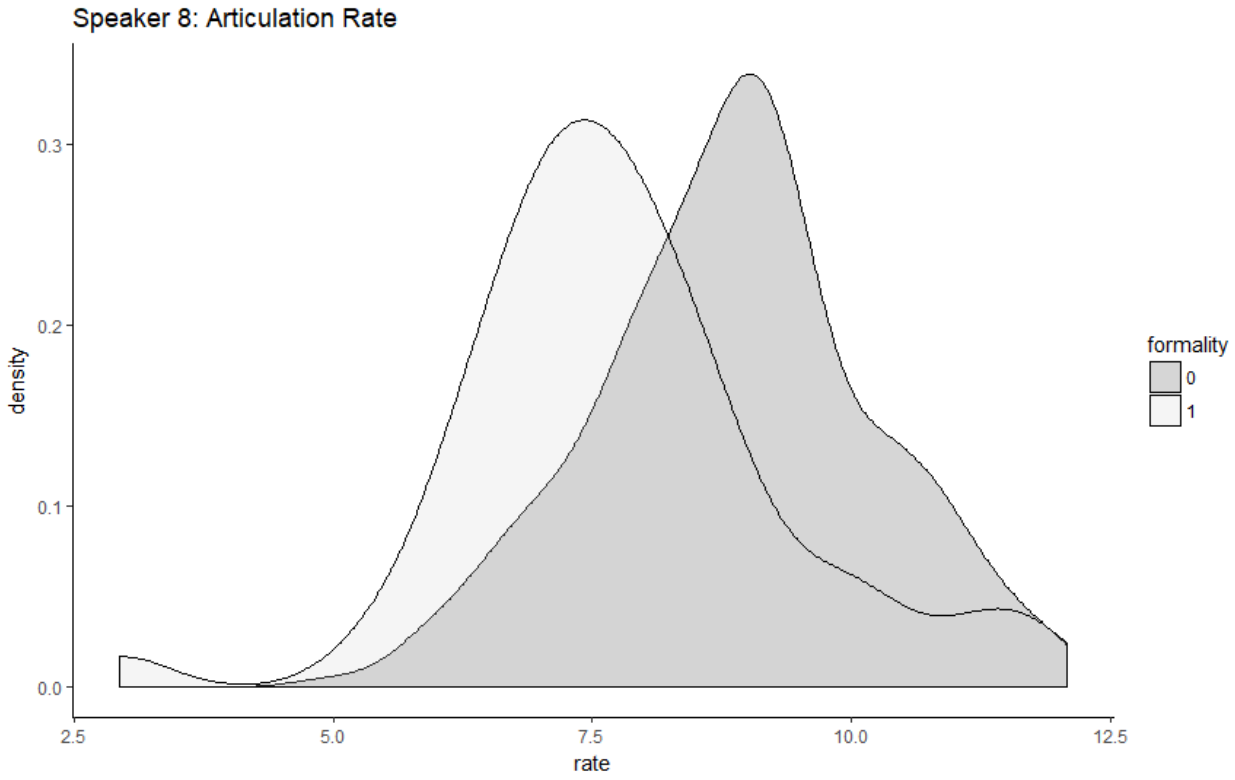

Speaker 7: f0
density
0.0
2.5
5.0
7.5
10.0
2.20
2.25
2.30
2.35
2.40
logf0
formality
0
1

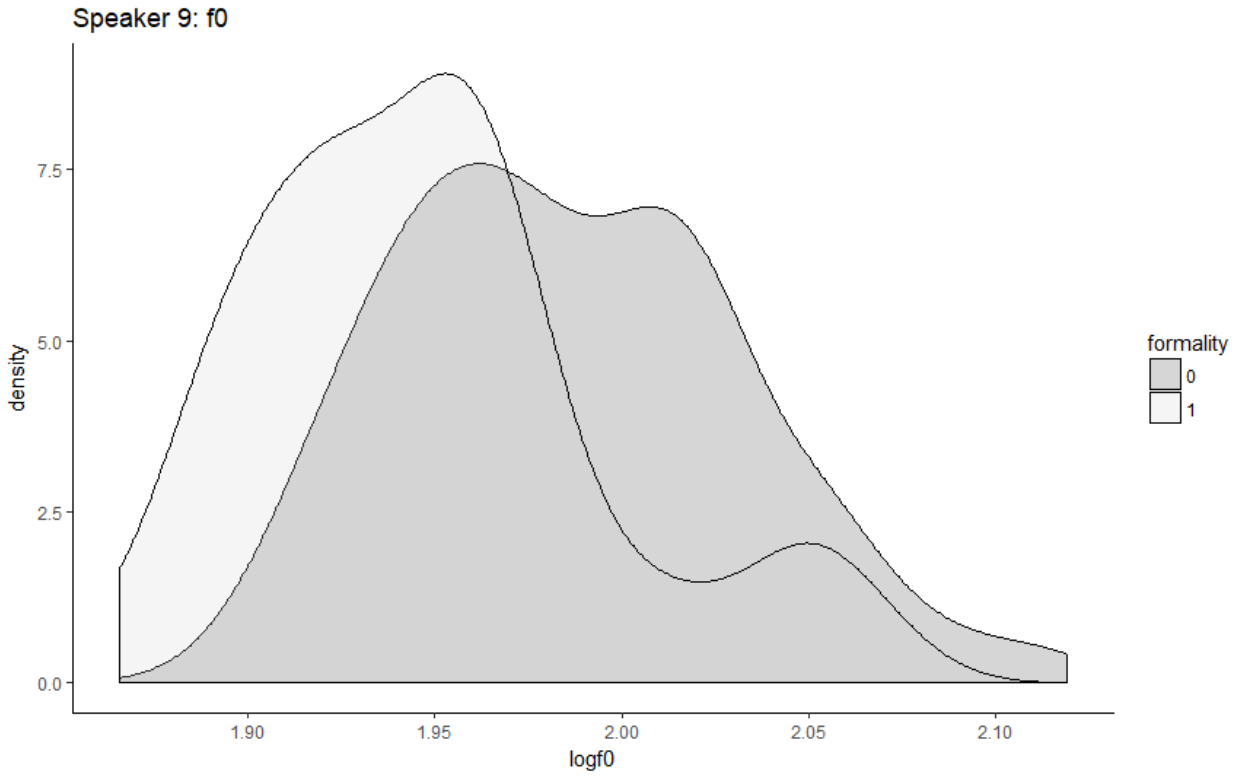

Speaker 7: f0 Range
density
0.010
0.005
0.000
0
100
200
range
formality
0
1

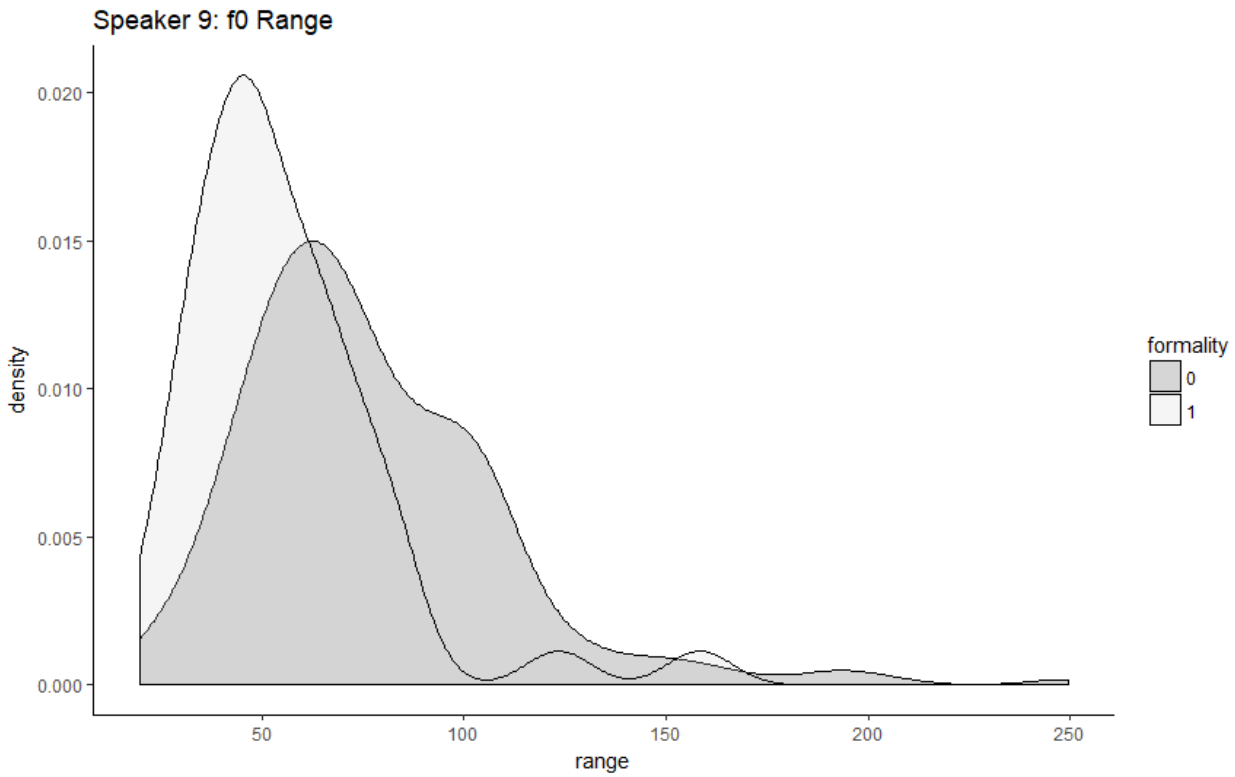

Speaker 7: Articulation Rate
density
0.3
0.2
0.1
0.0
5.0
7.5
10.0
12.5
rate
formality
0
1

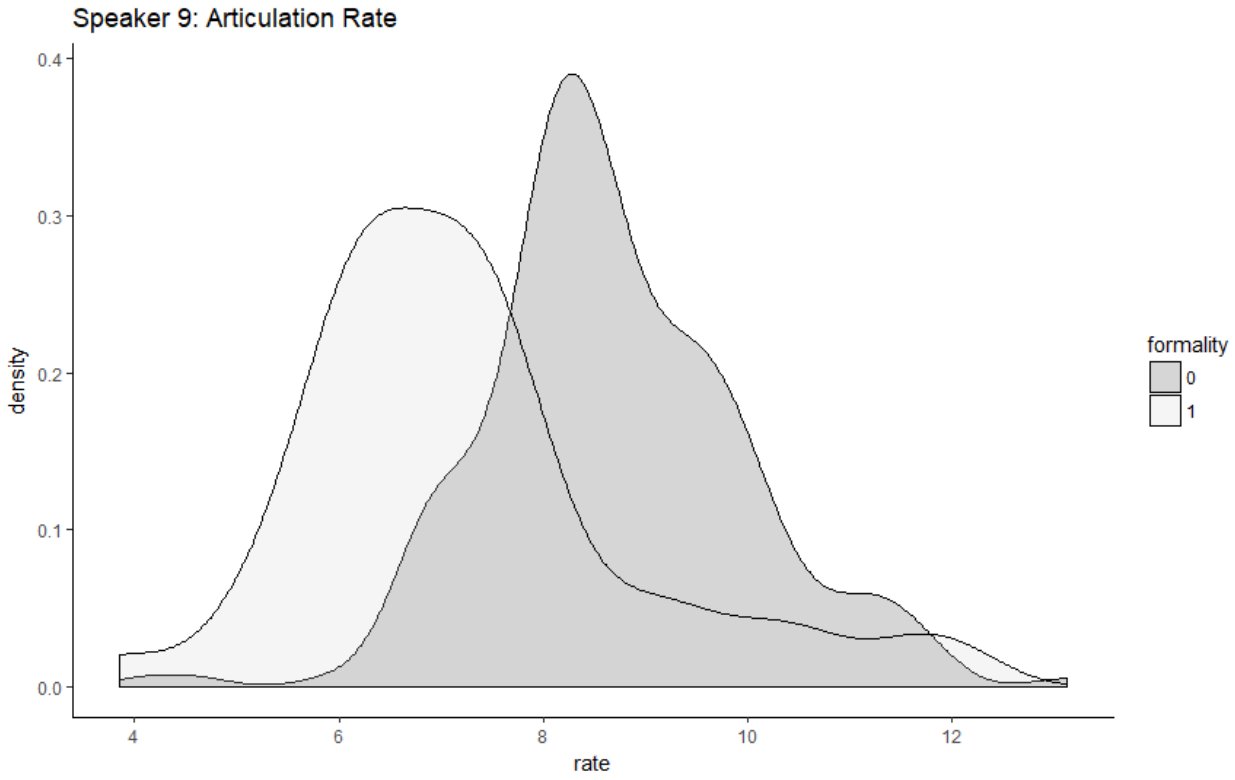

Speaker 8: f0
density
9
6
3
0
2.2
2.3
2.4
2.5
logf0
formality
0
1

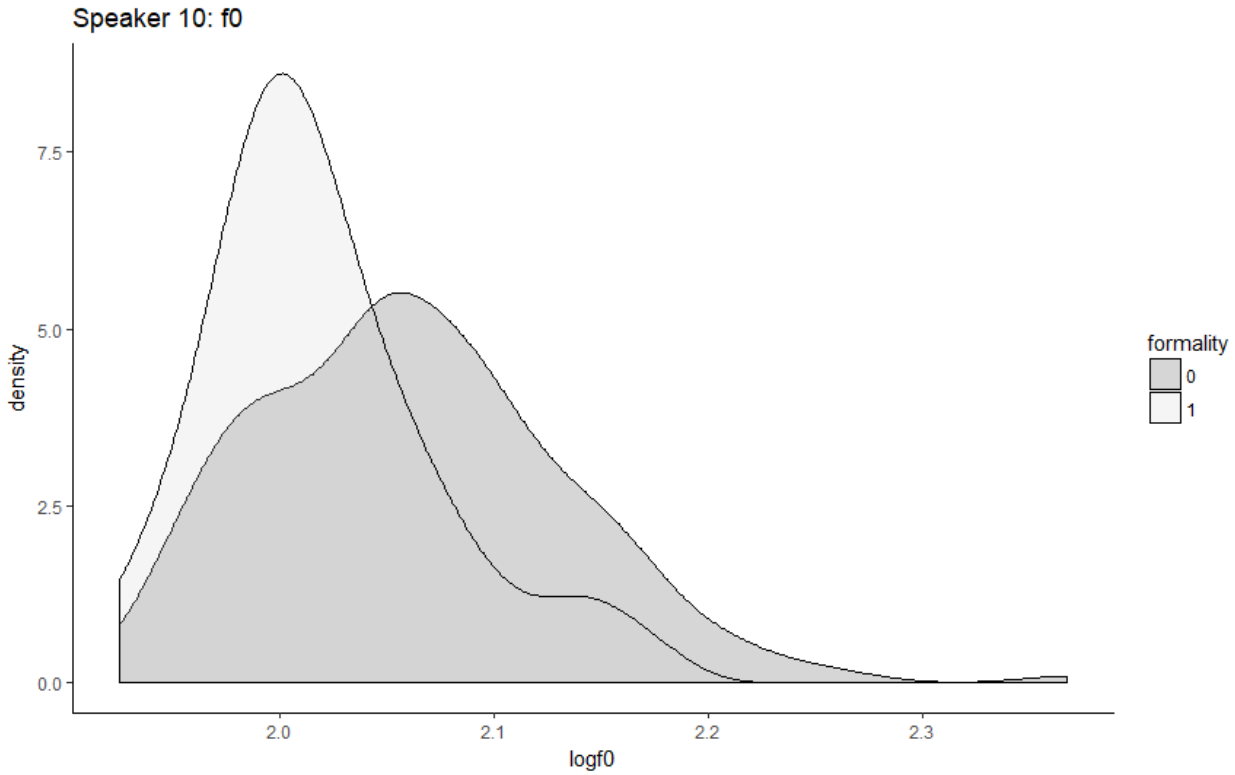

Speaker 8: f0 Range
density
0.0100
0.0075
0.0050
0.0025
0.0000
0
100
200
300
400
range
formality
0
1

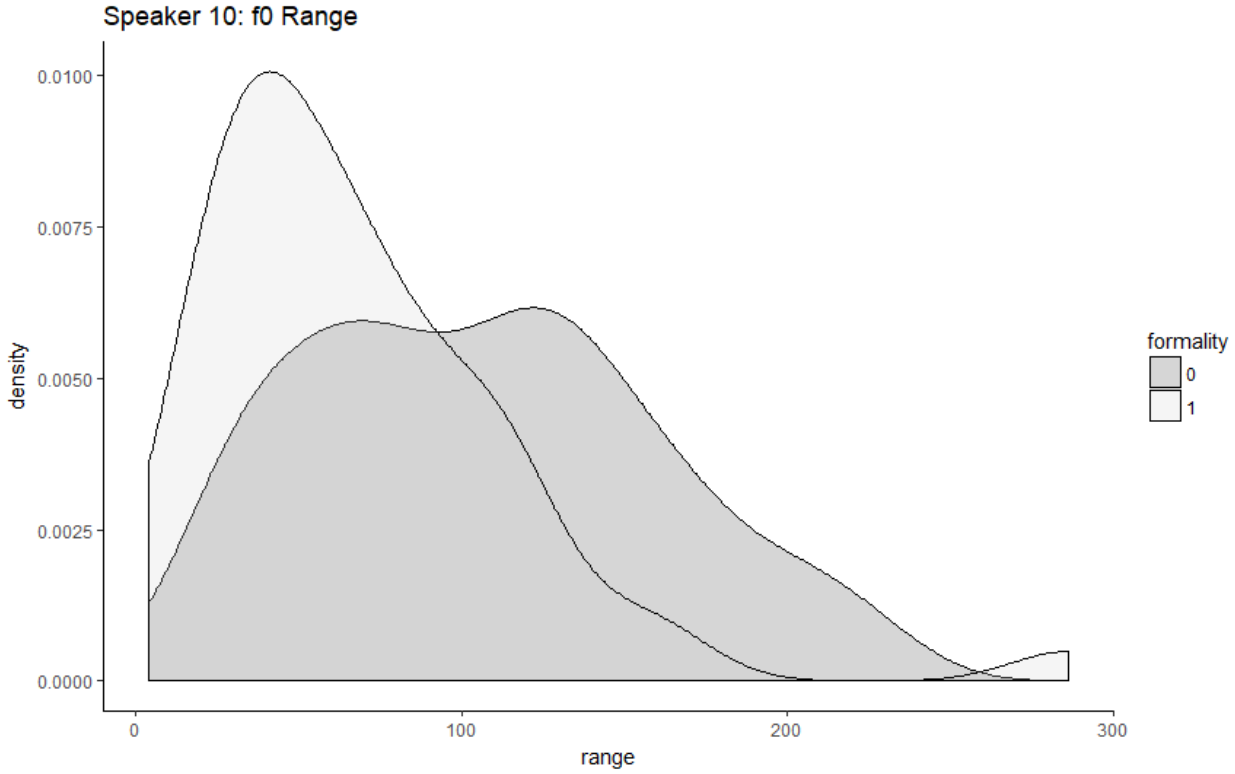

Speaker 8: Articulation Rate
density
0.0
0.1
0.2
0.3
2.5
5.0
7.5
10.0
12.5
rate
formality
0
1

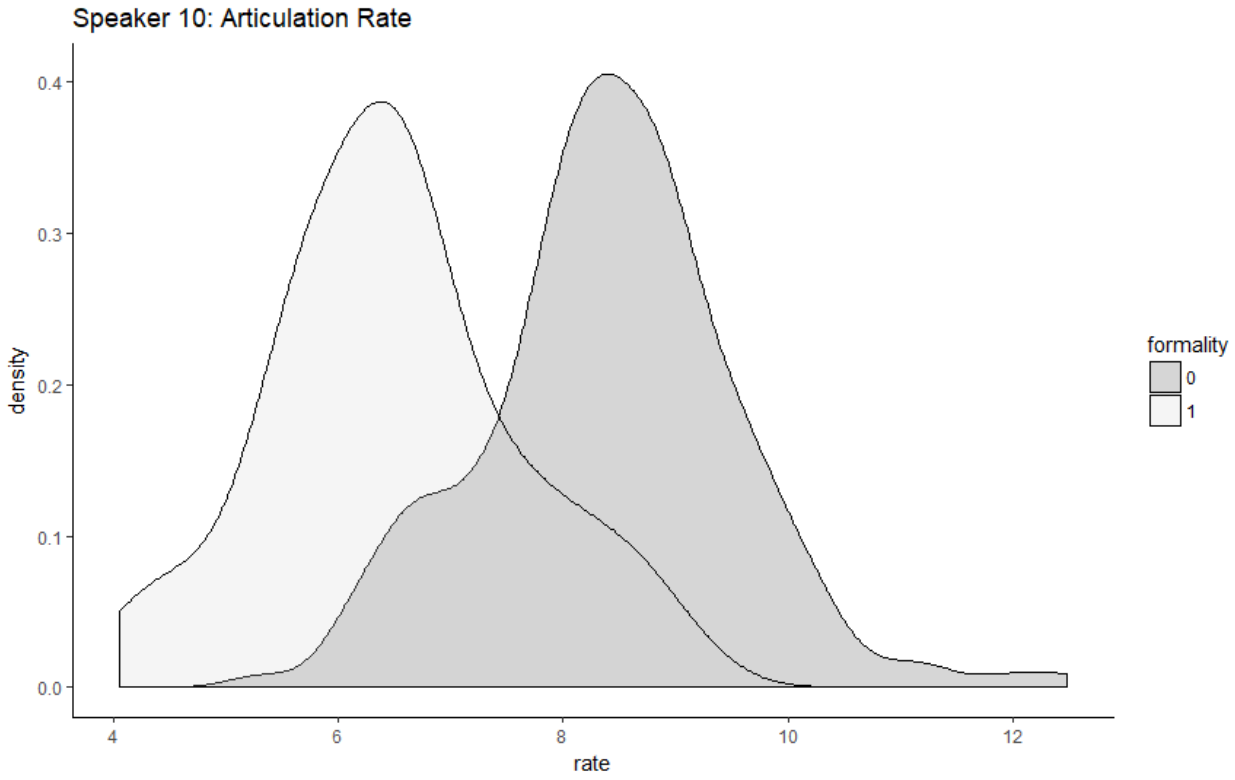

Speaker 9: f0
density
0.0
2.5
5.0
7.5
1.90
1.95
2.00
2.05
2.10
logf0
formality
0
1


Speaker 9: f0 Range
density
0.020
0.015
0.010
0.005
0.000
50
100
150
200
250
range
formality
0
1


Speaker 9: Articulation Rate
density
0.4
0.3
0.2
0.1
0.0
4
6
8
10
12
rate
formality
0
1


Speaker 10: f0
7.5
5.0
2.5
0.0
density
2.0
2.1
2.2
2.3
logf0
formality
0
1


Speaker 10: f0 Range
0.0100
0.0075
0.0050
0.0025
0.0000
density
0
100
200
300
range
formality
0
1



Appendix III – List of Common Interview Topics

More Formal Topics

1. Self-introductions.
2. Subject's work or topic of study.
3. Cultural differences between countries.
4. World travel.
5. Subject's family.

Less Formal Topics

1. The subject's and interviewer's hometowns.
2. Favorite sports/teams.
3. Favorite movies or TV shows.
4. Japanese food.
5. Subject's hobbies or leisure activities.
6. Subject's club activities in high school or university.
7. Things to do in Tachikawa city.

Appendix IV – Klatt Synthesizer Parameters

F_N refers to the formant frequencies, while B_N refers to the bandwidth.

f_0: Variable

Amplitude of voicing: Variable

F_1 : 500, B_1 : 60 Hz

F_2 : 1500, B_2: 90 Hz

F_3 : 2500, B_3 : 200 Hz

F_4 : 3300, B_4 : 250 Hz

F_5 : 3750, B_5 : 200 Hz

F_6 : 4900, B_6 : 1000 Hz

Frequency of nasal zero: 0 Hz

Bandwidth of nasal zero: 0 Hz

Frequency of nasal pole: 250 Hz

Bandwidth of nasal pole: 100 Hz

Amplitude of aspiration: 0 dB

Open quotient of voicing: 60 (smoother voice quality)

Amplitude of turbulence: 40 dB (simulates breathy voice quality)

Spectral tilt: 0 dB

Amplitude of frication: 0

Spectral Skew: 0

$A_{1...N}$ … Amplitude of formants 1 to N: Variable

B_Np … bandwidth of the parallel branch: Same as $B_{1...N}$

Amplitude of parallel nasal formant: 15 dB

Amplitude of bypass frication: 0 dB

Amplitude of voicing for the parallel branch: Variable

Gain: 60

References

Abercrombie, D. (1967). *Elements of General Phonetics.* Edinburgh University Press.

Abramson, A. S. (1987). Word-initial consonant length in Pattani Malay. *The 11th International Congress of Phonetic Sciences*. Tallinn, Academy of Sciences of the Estonian S.S.R. 6, 68–70.

Agaath, M., Sluitjer, C., & Van Heuven, V. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America.* 100(4), 2471-2485.

Álvarez, A., & Blondet, M. A. (2003). Cortesía y prosodia: un estudio de la frase cortés en el español de Mérida (Venezuela). In Z. E. Herrera & P. M. Butragueño (Eds.), *La tonía: Dimensiones fonéticas y fonológicas* (pp. 319- 330). México D. F El Colegio de México

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., & Sotillo, C. (1991). The HCRC map task corpus. *Language and Speech*, 34(4), 351-366.

Aono, M., Ichikawa, A., Koiso, H., Sato, S., Naka, M., Tutiya, S., & Suzuki, H. (1994). The Japanese map task corpus: an interim report. *Spoken language understanding and discourse processing, Japanese Society for Artificial Intelligence*, 25-30.

Arai, T. (1999). A case study of spontaneous speech in Japanese. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, 615-618. Berkeley, CA: Department of Linguistics, University of California.

Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE transactions on pattern analysis and machine intelligence,* (2), 179-190.

Baker, J. (1975). The DRAGON system--*An overview. IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1), 24-29.

Bard, E.G., Robertson D., & Sorace, A., (1996). Magnitude Estimation of Linguistic Acceptability, *Language*, 72, 32-68.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software.*

Bayes, T. and Price, R. (1763). An Essay Towards Solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M.A. and F.R.S. *Philosophical Transactions of the Royal Statistical Society of London*, 53, p.370–418.

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology, 3*(1), 255-309.

Boersma, P. & D. Weenink. (2017). Praat: doing phonetics by computer. http://www.fon.hum.uva.nl/praat/

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution, 24*(3), 127-135.

Box, G. E., & Tiao, G. C. (2011). *Bayesian inference in statistical analysis (Vol. 40).* John Wiley & Sons.

Brown, L., Winter, B., Idemaru, K., & Grawunder, S. (2014). Phonetics and politeness: Perceiving Korean honorific and non-honorific speech through phonetic cues. *Journal of Pragmatics*, *66*, 45-60.

Brown, P. & S. Levinson. (1987). *Politeness, Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Bybee, J. (2001). *Phonology and Language Use*. Cambridge: Cambridge University Press.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language* 82, 711–733.

Casella, G., & Berger, R. R. 2001*, Statistical Inference.* Duxbury Press.

Christensen, R. H. B. (2015). ordinal - Regression Models for Ordinal Data. R package version 2015.6-28. http://www.cran.r-project.org/package＝ordinal/.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues*. Cognition, 108(3), 804-809.*

Collier, R., & 't Hart, J. (1975). The role of intonation in speech perception, in: A. Cohen and S.G. Nooteboom (Eds.) *Structure and Process in Speech Perception*, Springer Verlag Heidelberg, 107-123.

Cook, H. M. (1998). Situational meanings of Japanese social deixis: The mixed use of the masu and plain forms. *Journal of Linguistic Anthropology 8(1),* 87-110.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. (1952). Some experiments on the perception of synthetic speech sounds. *The Journal of the Acoustical Society of America*, 24(6), 597-606.

Darwin, C. J. (1975). On the dynamic use of prosody in speech perception. *Status Report on Speech Research*, 42/43. Haskins Laboratories.

Dellwo, V. (2008). The role of speech rate in perceiving speech rhythm. *Speech Prosody* 2008 4(8), 375-378.

Den, Y. (2014). Chiba 3-way conversation corpus. Chiba University.

Eaton, J., Bateman, D., Hauberg, S., & Wehbring, R. (2015). GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations. URL: http://www.gnu.org/software/octave/doc/interpreter

Entropic Speech. Inc. (1989). *Entropic signal procesing systems*. Washington, DC: Author.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of human genetics*, 7(2), 179-188.

Fitzgerald, W. J., & Rayner, P. J. W. (1999). Bayesian signal processing. In *IEE Colloquium (December Digest),* 35-40.

Flanagan, J. L., & Saslow, M. G. (1958). Pitch discrimination for synthetic vowels. *The Journal of the Acoustical Society of America*, 30(5): 435-442.

Freeberg, T.M., & Lucas, J.R. (2009). Pseudoreplication is (still) a problem. *Journal of Comparative Psychology*, 123, 450-451.

Fujisaki, H., & Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan (E), 5*(4), 233-242.

Fukui, N. (1986). Leftward Spread: Compensatory lengthening and gemination in Japanese. *Linguistic Inquiry*. 17(2), 359-364.

Gahl, S. (2008). Time and thyme are not homophones: the effect of lemma frequency on word durations in spontaneous speech. *Language* 34(3), 474-496.

Geisler, W. S. (2003). Ideal observer analysis. *The visual neurosciences*, *10*(7), 12-12.

Goel, V., & Byrne, W. J. (2000). Minimum Bayes-risk automatic speech recognition. *Computer Speech & Language*, *14*(2), 115-135.

Grabe, E. & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis, in C. Gussenhoven and N. Warner (eds.) *Papers in Laboratory Phonology 7*, Berlin, New York: Mouton de Gruyter.

Grabe, E., Kochanski, G., & Coleman, J. (2007). Connecting intonation labels to mathematical descriptions of fundamental frequency. *Language and Speech* 50(3), 281-310.

Grawunder, S., Oertel, M., & Schwarze, C. (2014). Politeness, culture, and speaking task – Paralinguistic prosodic behavior of speakers from Austria and Germany. *Proceedings of the international conference on Speech Prosody*, Dublin, Ireland 159–163.

Guion, S. (1995). Word frequency effects among homonyms. *Texas Linguistic Forum* 35, 103–116.

Guion, S. & Idemaru, K. (2008). Acoustic covariants of length contrast in Japanese stops. *Journal of the IPA*. 38, 167-286.

Halliday, M.A.K. & R. Hasan. (1976). *Cohesion in English*, London: Longman.

Harrington, J., Palethorpe, S., & Watson, C.J. (2007). Age-related changes in fundamental frequency and formants: a longitudal study of four speakers. *Interspeech 2007,* 2753-2756.

Hawkins, J. (2006) Gradedness as relattive efficiency in the processing of syntax and semantics. In Fanselow, G., C. Fery, R. Vogel, & M. Schlesewsky eds. (2006). *Gradience in Grammar*. Oxford: Oxford University Press.

Hart, J. T., Collier, R., & Cohen, A. (2006). *A perceptual study of intonation: an experimental-phonetic approach to speech melody*. Cambridge University Press.

Henton, C. G. (1989). Fact and fiction in the description of female and male pitch. *Language & Communication, 9*(4), 299-311.

Henton, C. G. (1995). Pitch dynamism in female and male speech. *Language &*

Communication, 15(1), 43-61.

Hidalgo Navarro, A., & Cabedo Nebot, A. (2014). On the importance of the prosodic component in the expression of linguistic im/politeness. *Journal of Politeness Research.* 10(1), 5-27.

Hinds, J. (1976). *Japanese Discourse Structure.* Tokyo: Kaitakusha.

Hinds, J. (1978). Anaphora in Japanese conversation. *Anaphora in discourse*, 136-179.

Hiramoto, M. (2010). Untterance final position and projection of femininity in Japanese. *Gender and Language*: 4(1), 99-124.

Hori, M. (1986). A sociolinguistic analysis of Japanese honorifics. *Journal of Pragmatics* 10, 373-386.

Hübscher, I., Borràs-Comes, J., & Prieto, P. (2017). Prosodic mitigation characterizes Catalan formal speech: The Frequency Code reassessed. *Journal of Phonetics*, 65, 145-159.

Ide, S. (1982). Japanese sociolinguistics politeness and women's language. *Lingua* 57, 357-385.

Iles, J., & Ing-Simmons, N. (1994). Klatt: A Klatt-style speech synthesizer implemented in C (Version 3.0.4) [computer software]. *CMU Artificial Intelligence Repository.*

Inoue, M. (2002). Gender, Language, and Modernity: towards an effective history of Japanese women's language. *American Ethnologist* 29(2): 392-422.

Ito, M. (2001). Rating experiments of spoken Japanese politeness. In *Proceedings of the*

Postgraduate Conference, University of Edinburgh.

Ito, M. (2002). Japanese politeness and supersegmentals – a study based on Natural Speech Materials. *Speech Prosody 2002.*

Jones, K. & T. Ono. (2008). *Style Shifting in Japanese.* John Benjamin's: Philadelphia.

Kawahara, S. (2006). A faithfulness ranking projected from a perceptibility scale: The case of [+voice] in Japanese. *Language* 82(3), 536-574.

Kawahara, S. (2015). The phonetics of obstruent geminates, sokuon. *The Mouton Handbook of Japanese Language and Linguistics. Berlin, Germany: Mouton de Gruyter*.

Kay, S. M. (1993). Statistical signal processing. *Estimation Theory*, *1*.

Klatt, D. H. (1973). Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. *The Journal of the Acoustical Society of America*, *53*(1), 8-16.

Klatt, D. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America* 67(3): 971-995.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, *87*(2), 820-857.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: recognize the familiar,

generalize to the similar, and adapt to the novel. *Psychological review*, 122(2): 148–203.

Kubozono, H. (1993). *The organization of Japanese prosody*. Tokyo: Kurosio.

Kubozono, H. (2011). Japanese pitch accent. *The Blackwell companion to phonology*, *5*, 2879-2907.

Kubozono, H. (2012). Varieties of pitch accent systems in Japanese. *Lingua* 122(13): 1395-1414.

Kubozono, H. (Ed.). (2015). *Handbook of Japanese Phonetics and Phonology* (Vol. 2). Walter de Gruyter GmbH & Co KG.

Laan, P. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of spontaneous and read speaking style. *Speech Communication* 27: 43-65.

Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press.

de Laplace, P. S. (1820). *Théorie analytique des probabilités* (Vol. 7). Courcier.

Lazic, S.E. (2010). The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neuroscience*, 11, 1-17.

Lin, H., Kwock-Ping, J. T., & Fon, J. (2006). An acoustic study on the paralinguistic prosody in the politeness talk in Taiwan Mandarin. *Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics*, Athens, Greece, 173-176.

Lisker L, Abrahmson AS. (1964). Cross-language study of voicing in initial stops. *Word,* 20, 384–422.

Local, J., & Simpson, A. (1999). Phonetic implementation of geminates in Malayalam nouns. *Dept. of Language and Linguistic Science, University of York, U.K.*

Loveday, L. (1981). Pitch, politeness and sexual role: An exploratory investigation into the pitch correlates of English and Japanese politeness formulae. *Language and Speech*, *24*(1), 71-89.

Loveday, L. (1986). *Explorations in Japanese sociolinguistics*. John Benjamins Publishing.

Maekawa, K. (2003). Corpus of spontaneous Japanese: Its design and evaluation. *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003.*

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., ... & Den, Y. (2014). Balanced corpus of contemporary written Japanese*. Language Resources and Evaluation*, 48(2), 345-371.

Matsumoto, Y. (1988). Rexamination of the universality of face: Politeness phenomena in Japanese. *Journal of Pragmatics* 12: 403-426.

McCarthy, J. J. (1979) Formal Problems in Semitic Phonology and Morphology. Doctoral dissertation, MIT, Cambridge, Massachusetts.

Mester, A. & J. Ito. (1995). Japanese phonology. In Goldsmith, J. *The Handbook of*

Phonological Theory. Blackwell Publishers. 817-838.

Minegishi-Cook, G. (2008). *Socializing Identities Through Speech Style: Learners of Japanese as a Foreign Language.* Multilingual Matters: Bristol.

Morley, E., Klabbers, E., van Santen, J. P., Kain, A., & Mohammadi, S. H. (2012). Synthetic F0 Can Effectively Convey Speaker ID in Delexicalized Speech. *INTERSPEECH 2012*: 434-437.

Moulines, E., Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 9 (5/6), 453-467.

Munson, B., Johnson, J., & Edwards, J. (2011). The role of clinical experience in speech-language pathologists' perception of subphonemic detail in children's speech. *American Journal of Speech-Language Pathology,* 21(2), 124.

Mwangi, S., Spiegl, W., Honig, F., Haderlein, T., Maier, A., & Noth, E. (2009). Effects of vocal aging on fundamental frequency and formants. *Dutch Acoustical Society/German Acoustical Society 2009*: 1761-1764.

Nakamura, K., Dehaene, S., Jobert, A., Le Bihan, D., & S. Kouider. (2005). Subliminal convergence of kanji and kana words: further evidence for functional parcellation of the posterior temporal cortex in visual word perception. *Journal of Cognitive Neuroscience* 17(6), 954-968.

Nakamura, M., Iwano, K., & Furui, S. (2007) Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech and Language* 22, 171-184.

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas—Part II. *Language and Linguistics Compass*, 10(11), 591-613.

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of language and social psychology*, 18(1): 62-85.

Nolan, F. (2003). Intonational equivalence: an experimental evaluation of pitch scales. In *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona* (Vol. 39).

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological review*, 115(2), 357–395.

Ofuka, E., McKeown, J., Waterman, M., & Roach, P. (2000). Prosodic cues for rated politeness in Japanese speech. *Speech Communication* 32, 199-217.

Ogino, T. & M. Hong. (1992). Nihongo onsei no teineisa ni kansuru kenkyuu (A study on politeness in Japanese speech). In: Kunihiro, T. (Ed.), *Nihongo intonation no jittai to bunseki (The State-of-the-art and Analysis of Japanese Intonation).* pp. 215-258. Tokyo: Monbushou.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of

voice. *Phonetica*, 41, 1–16.

Ohara, Y. (2001). Finding one's voice in Japanese: A study of the pitch levels of L2 users. In: A. Pavlenko, A. Brackledge, I. Piller, & M. Teutsch-Dwyer (Eds.), *Multilingualism, second language learning, and gender* (pp. 231–254). New York: Mouton de Gruyter.

Ohara, Y. (2004). Prosody and gender in workplace interaction: exploring constrains and resources in the use of Japanese In Okamoto and JS Smith (Eds.) *Japanese Language, Gender, and Ideology: Cultural Models and Real People* 222-239. New York: Oxford University Press.

Okamoto, S. (1999). Situated politeness: manipulating honorific and non-honorific expressions in Japanese conversations. *Pragmatics*, 9(1), 51-74.

Pagel, V., Carbonell, N., & Laprie, Y. (1996). A new method for speech delexicalization, and its application to the perception of French prosody. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on* Vol. 2: 821-824. *IEEE.*

Payne, E. (2005). Phonetic variation in Italian consonant gemination. *Journal of the IPA* 35(2), 153-181.

Pierrehumbert, J., & Beckman, M. (1988). Japanese tone structure. *Linguistic inquiry monographs*, (15), 1-282.

Pisoni, D. B. (1997). Perception of synthetic speech. In *Progress in speech synthesis* (pp. 541-560). Springer New York.

Pizziconi, B. (2002). Re-examining Japanese politeness, face, and the Japanese language. *Journal of Pragmatics* 35(2003), 1471-1506.

Prince, A. & P. Smolensky. (2002). *Optimality Theory: Constraint Interaction in Generative Grammar.* Wiley.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL: *https://www.R-project.org/.*

Ramsay, J. O. (2006). *Functional data analysis*. John Wiley & Sons, Inc.

Ridouane, R. (2007). Gemination in Tashlhiyt Berber: an acoustic and articulatory study. Journal of the International Phonetic Association, 37(2), 119.

Sagisaka, M., & M. Miyatake. (1988) Prosodic characteristics and their control in Japanese speech under varying speech styles. *Journal of the ASA Supplement 1,* 83, S27.

Schütze, C. T., & Sprouse, J. (2014). Judgment data. in Podesva, Robert J., and Devyani Sharma, eds. *Research methods in linguistics,* pp. 27-50. Cambridge University Press, 2014.

Siegal, M. & S. Okamoto. (2003). Toward reconceptualizing the teaching and learning of gendered speech styles in Japanese as a Foreign Language. *Japanese Language and Literature* 37 (1): 49–66.

Slowiaczek, L. M., & Nusbaum, H. C. (1985). Effects of speech rate and pitch contour on the perception of synthetic speech. *Human Factors: The Journal of the Human Factors*

and Ergonomics Society, 27(6): 701-712.

Smith, C. (1993). Prosodic patterns in the coordination of vowel and consonant gestures. *Haskins Laboratory Report On Speech Research*. 115/116, 45-55.

Sreetharan, C. (2004). Students, *sarariiman* (pl.), and seniors: Japanese men's use of the 'manly' speech register. *Language in Society* 33: 81-107.

Stan Development Team (2016). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.13.1. http://mc-stan.org/.

Stevens, S. S., & Volkmann, J. (1940). The relation of pitch to frequency: A revised scale. *The American Journal of Psychology, 53*(3), 329-353.

Studdert-Kennedy, M. (1979). Speech Perception. *Status Report on Speech Research* 59/60. Haskins Laboratories.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *The Journal of the Acoustical Society of America, 69*(5), 1465-1475.

Tsuji, A. (2004). The case study of high pitch register in English and in Japanese: Does high pitch register relate to politeness? *Seijo English Monographs*, 37, 227‐260.

van Santen, J., & Mobius, B. (2000). A quantitative model of F_0 generation and alignment. In A. Botinis, Ed. *Intonation – Analysis, Modelling and Technology.* Kluwer academic publishers, 269–288.

Venditti, J. J., Jun, S. A., & Beckman, M. E. (2014). Structures in Japanese, Korean, and

English. *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, 287-311.

Wang, X., Chen, R., & Liu, J. S. (2002). Monte Carlo Bayesian signal processing for wireless communications. *The Journal of VLSI Signal Processing*, *30*(1), 89-105.

Warner, N., & Arai, T. (2000). Japanese mora-timing: A review. *Phonetica*, 58(1-2), 1-25.

Winter, B. (2011). Pseudoreplication in phonetic research. *Proceedings of the International Congress of Phonetic Science,* 2137-2140. Hong Kong, August 2011.

Winter, B., & Grawunder, S. (2012). The phonetic profile of Korean formal and informal speech registers. *Journal of Phonetics,* 40, 808-815.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499.