

ATTRIBUTE BASED SHARED HIDDEN LAYERS FOR CROSS-LANGUAGE KNOWLEDGE TRANSFER

Vipul Arora, Aditi Lahiri

Faculty of Linguistics, Philology and Phonetics,
University of Oxford, U.K.
{vipul.arora, aditi.lahiri}@ling-phil.ox.ac.uk

Henning Reetz

Goethe University,
Frankfurt am Main, Germany
reetz@em.uni-frankfurt.de

ABSTRACT

Deep neural network (DNN) acoustic models can be adapted to under-resourced languages by transferring the hidden layers. An analogous transfer problem is popular as few-shot learning to recognise scantily seen objects based on their meaningful attributes. In similar way, this paper proposes a principled way to represent the hidden layers of DNN in terms of attributes shared across languages. The diverse phoneme sets of different languages can be represented in terms of phonological features that are shared by them. The DNN layers estimating these features could then be transferred in a meaningful and reliable way. Here, we evaluate model transfer from English to German, by comparing the proposed method with other popular methods on the task of phoneme recognition. Experimental results support that apart from providing interpretability to the DNN acoustic models, the proposed framework provides efficient means for their speedy adaptation to different languages, even in the face of scanty adaptation data.

Index Terms— Deep neural networks adaptation, knowledge transfer, cross-lingual ASR, phonological features, zero-shot learning

1. INTRODUCTION

Cross-language model transfer is an important area of research for automatic speech recognition (ASR), where acoustic models trained over resource-rich languages are adapted to under-resourced languages. Apart from having different lexicons and grammars, different languages mostly have different phoneme sets with different number of phonemes; adding to the challenge involved in model transfer across languages. The main questions in transfer learning are what and how to transfer [1]. Many diverse solutions have been proposed to these two questions, using different kinds of acoustic models. Some [2] have proposed to use a global set of phonemes which are shared by different languages; hence, transferring phoneme-recognisers across languages. Most methods

transfer more abstract representations in the form of model parameters.

This paper focuses on deep neural network (DNN) based acoustic models, which are presently the state of the art. They provide high accuracy for ASR, but require large amount of training data, thereby posing a challenge for under-resourced languages. The weights and biases of the hidden layers of DNN store the information learned from the training data. They are hard to interpret and hence, their transfer across different domains (e.g. languages) is quite challenging. Nevertheless, towards this end, popular approaches use either tandem or hybrid configuration. In tandem configuration, DNN layers are trained to extract features called bottleneck (BN) features. These BN features [3, 4, 5] are then used as input for a language specific ASR system, which can be a conventional hidden Markov model (HMM) with Gaussian mixture models (GMMs) as acoustic models. In some works [6], the extracted BN features are appended to conventional acoustic parameters like mel-frequency cepstral coefficients (MFCCs), etc. In hybrid configuration, DNNs are trained to directly estimate the state probabilities for HMM. Many methods [7, 8] transfer the lower layers (closer to input side), and re-train the higher level layers afresh over the training data of the target domain. Some works [9] also use data from different languages for pretraining the DNN layers in an unsupervised fashion.

Another paradigm to transfer models to unseen domains is known as zero-shot learning (or few-shot learning if a little adaptation data is available). It is popular, for example, in the area of computer vision to classify instances of unseen objects, i.e. objects not present in training data [10, 11, 12]. The idea is to represent every object in terms of meaningful attributes which are shared by even unseen objects. For instance, a zebra is represented with attributes like striped and four-legged. Models, typically support vector machines, are trained to detect these attributes from training data, which does not have any instance of zebra, and hence can be used to detect a zebra.

In this paper, we use this idea of zero-shot learning to achieve a better control over DNN layers to be transferred.

This work was supported by the ERC Proof of Concept FLEX-SR award no. 632226.

However, we have also got some training data from the target language to adapt the models. The phonemes are represented in terms of attributes which are shared by phonemes across different languages. The detector of these attributes can be then be transferred to low-resource languages. This provides two immediate advantages - first, we obtain a principled way to transfer DNN layers across languages, and second, even the shared layers can be re-trained over the target language in a reliable and efficient way.

We use phonological features as the meaningful attributes shared across languages. These features have been popular previously for speech recognition. They are found to be a useful representation of speech with diverse applications, both in speech recognition as well as synthesis. King and Taylor [13] use phonological features, based on three different phonological models, as input to GMM-HMM phoneme recogniser, and obtain a performance similar to that obtained using cepstral features as the input. Scanlon *et al.* [14] propose to divide all phonemes into phonological classes and use modular classifiers for phonemes within the same phonological class, achieving better performance in phoneme recognition. Siniscalchi *et al.* [15] show that phonological features bring improvement in performance for large vocabulary continuous speech recognition. Jansen and Niyogi [16] detect such features as acoustic events, and use point process based system for keyword spotting. Cernak *et al.* [17] propose using phonological features for speech vocoding. Nagamine *et al.* [18] show phonological feature specific activation of neurons of a DNN trained for phoneme recognition.

The use of phonological features for cross language ASR has also been considered. Çetin *et al.* [19] use articulatory features to derive bottleneck features with shallow neural networks and use them for cross-language ASR. Stuker *et al.* [20] use multi-lingual articulatory features. Siniscalchi *et al.* [21] use shallow neural networks for detection of features and, subsequently from them, HMM state probabilities. However, we use phonological features as a way for cross-language transfer of hidden layers of DNN.

2. EXTRACTION OF SHARED ATTRIBUTES

Phonological features form natural classes which are shared across the world's languages. Hence, they are an apt candidate to serve as meaningful attributes for model transfer. Each phoneme is characterised by a group of phonological features based on the manner and place of articulation. Thus, a single feature can characterise a set of phonemes. For the present work, we use a set of 18 such features, to characterise different phonemes in English and German. A silence feature (SIL) is added to make the total number of attributes to be 19.

The features used here [22] include vocalic (VOC) for vowels and consonantal (CONS) for consonants. Pitched phonemes have the feature sonorant (SON). Features low (LOW) and high (HIGH) correspond to the tongue height;

while features coronal (COR) and dorsal (DOR) pertain to tongue's frontness or backness; and labial (LAB) indicates the involvement of lips. Short vowels have the retracted tongue root (RTR) feature. Feature continuant (CONT) marks the consonants continuing in time while feature stop (STOP) denotes the stop consonants. A voiced consonant carries voicing feature (VOICE) to distinguish from its unvoiced counterpart. Any consonant involving obstruction of the vocal tract is associated with the feature obstruent (OBS), with consonants that involve turbulent streaming of air having the feature strident (STR). Nasal sounds carry the feature nasal (NAS). Features like rhotic (RHO), lateral (LAT) and radical (RAD) denote certain kinds of consonants, namely r-like, l-like and h-like, respectively. Silence (SIL) marks pauses as well as closures in stops.

2.1. DNNs for Attribute Extraction

The input to the DNN are the acoustic parameters (APs) extracted from speech signal. Short-time FFT power spectra are calculated with 25ms window and 10ms hop size. They are binned with 23 Mel-scaled filters, whose log scaled outputs form the APs for each time frame. Mean and variance normalisation is applied to these APs, to be used as input to DNN with a context of ± 5 frames.

For attribute extraction, the DNN consists of three hidden layers, with 500, 500 and 200 neurons, respectively, each with ReLU (rectified linear unit) non-linearity. The output layer corresponds to the 19 attributes, with sigmoid activation function at each neuron.

The target values of attribute extraction DNNs are determined by canonical representation of the phonemes in terms of their phonological features. Phoneme state level segmentation is obtained by forced alignment of training data with tri-state GMM-HMM based phoneme recogniser using 13 MFCCs with delta and delta-delta appended. Although the onset and offset times of phonological features do not synchronise exactly with those of phonemes, the DNNs are able to take care of asynchrony [13]. The target values are set to binary $\{0, 1\}$ values denoting the absence and presence of the corresponding attributes at a particular time frame.

Training is performed to minimise the squared-error objective function. Network weights are updated using mini-batch based stochastic gradient descent algorithm with Nesterov momentum. The learning rate decays linearly from 10^{-2} to 10^{-5} , and the minibatch size increases linearly from 256 to 1024, with each epoch. DNNs are implemented using Lasagne library (github.com/Lasagne).

3. RECOGNITION SYSTEM

The recognition system is based on DNN-HMM framework [1]. For mapping from attribute space to phonemes, the neural network consists of single hidden layer with 500 neurons and

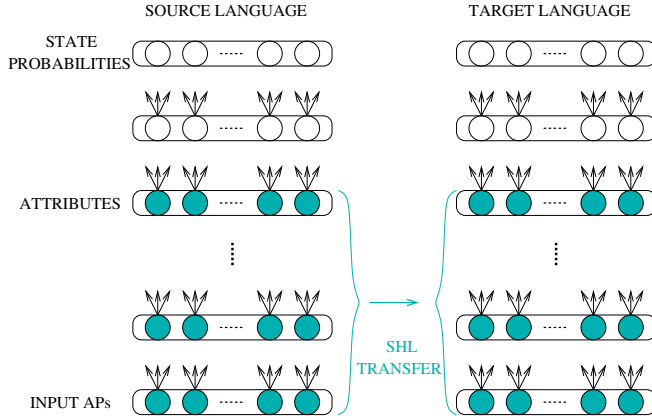


Fig. 1: Schematic for cross language transfer of attribute-based shared hidden layers (SHL) of DNN.

ReLU activation function. The attributes are fed directly into this network, without any transformation or stacking.

Output neurons correspond to the conditional state posterior probabilities of monophone states of HMM, numbering 3 per phoneme. Monophone models are used here, instead of triphone ones, as some of our preliminary experiments have indicated that they have better portability than triphone models. However, we leave the detailed exploration in this direction for future.

The target values are obtained using the forced alignment explained in Sec. 2.1. Output layer has softmax non-linearity. Training is performed to minimise categorical cross-entropy objective function. The weight update scheme is the same as those for attribute extraction DNNs (Sec. 2.1). The network outputs represent posterior probability of states. These are converted into likelihoods, by dividing with prior probabilities, to be used by HMM. The transition probabilities come from the original GMM-HMM system used for alignment. Bigram language model is trained over phoneme sequences in training data and is used for decoding. The decoding system is implemented using the Kaldi toolkit [23].

4. EXPERIMENTAL SETUP

4.1. Datasets

For training and evaluation, we use English as the source language and German as the target language. Apart from other differences in phonemes, German has some coronal (COR) vowels, carrying an additional labial (LAB) feature, not present in English. We use the TIMIT speech corpus¹ for English and the KIEL read speech corpus² for German. They consist of clean speech recorded in studio settings. However, we corrupt them with additive white Gaussian noise

Table 1: Attributes associated with phonemes of English (TIMIT) and German (KIEL)

Attribute	English	German
VOC	ah axr ax ax-h aa ae eh ih ix iy ao uh ux uw er	6 @r 2: 9 @ a a: e E e: E: I i: o: O O: U u u: Y y:
CONS	b ch d dh dx el em en eng f g hh hv jh k l m n ng nx p r s sh t th v z zh	bh tS dh N f gh h h dZ kh l m n N ph r s S th v z Z C x
CONT	dh el f hh hv l s sh th v z zh	l f h h l s S v z Z C x
OBSTR	b ch d dh dx f g hh hv jh k p s sh t th v z zh	bh tS dh f gh h dZ kh ph s S th v z Z C x
STR	ch s sh t th z zh	tS s S th z Z
VOICE	b d dh dx g jh v z zh	bh dh gh dZ v z Z
SON	ah axr ax ax-h aa ae eh el em en eng ih ix hv iy ao l m n ng nx ao r w y uh ux uw er	6 @r 2: 9 @ a a: E e e: E: N I h i: o: O l m n N O: r j U u u: Y y:
STOP	b ch d dx g jh k p t	bh tS dh gh dZ kh ph th
LOW	aa ae	a a: E
HIGH	ch ih ix iy jh sh w y zh uh ux uw	tS I i: dZ S j Z U u u: Y y:
LAB	b em f ao m p v w uh ux uw	2: 9 bh m f o: O m O: ph v U u u: Y y:
COR	ae eh ch d dh dx eh el en ih ix iy jh l n nx r s sh t th y z zh	2: 9 E e tS dh E e: E: I i: dZ l n r s S th j z Z Y y:
DOR	aa eng g ao k ng w uh ux uw	a a: N gh o: O kh N O: C U u u: x
RTR	ah axr ax ax-h eh ih w uh er	6 @r 9 @ e E e: E: I U Y
NAS	em en eng m n ng nx	m n N
LAT	el l	l l
RHO	r	r
RAD	hh hv	h
SIL	bcl dcl epi gcl h kcl pau pcl q tcl	b d g k p t Q sil

with 20dB SNR. Both these data sets are labelled at phoneme level. We use the transcription to get only the sequence of phonemes, without the time-alignment information, which we generate using forced alignment.

TIMIT transcriptions are based on 61 phonemes and KIEL transcriptions are based on 64 phonemes. Phonemes associated with each phonological feature are shown in Tab. 1. Certain phonemes, called diphthongs, consist of two vowels, thereby exhibiting change of features over time. Such phonemes are not considered for training and evaluation of attribute detection, but are used for training and evaluating phoneme estimation. In TIMIT, phonemes ay, aw, ey, ix, ow, oy are diphthong vowels. In KIEL, all vowels with suffix ‘6’, and vowels aI, aU, eI, oU, OY are diphthongs. In German, certain consonants – pf and ts – consist of two consonants, and hence, are not used for training.

Experiments involve three stages – training over the source language, adaptation over the target language and testing over the target language. The training set of English sentences comes from 3608 files from the TIMIT corpus. Total duration of these files is 2 hours 50 minutes. We name this as the TIMIT-train set. For adaptation over German, we use different numbers of files from the KIEL corpus for different experiments. This adaptation set is referred to as the KIEL-adaptation set. A small adaptation set is chosen

¹catalog.ldc.upenn.edu/LDC93S1

²www.ipds.uni-kiel.de/publikationen/kcrsp.en.html

here so as to evaluate the applicability over limited resource languages. The test set consists of 896 files from the KIEL corpus, with a total duration of 1 hour 5 minutes. We call this as the KIEL-test set.

Apart from the above-mentioned three sets, another test set, namely the TIMIT-test, from the TIMIT corpus is used to test the efficacy of proposed attribute-based representation over the source language itself. This set consists of 192 files with a total duration of 10 minutes. The speaker sets of the all the above datasets are mutually non-overlapping.

4.2. Proposed Method

4.2.1. Training

The DNN is trained over the source language, using the TIMIT-train set to estimate the attributes as described in Sec. 2.1.

4.2.2. Adaptation

Adaptation is carried out over the target language, using the KIEL-adaptation set. It is done in two steps. First, the shared hidden layers (SHL) are adapted by retraining the DNN for estimation of attributes from the input APs. Weights are updated using the scheme described in Sec. 2.1. Second, the layers for estimation of phoneme states from the attributes is trained wholly using the KIEL-adaptation set. This training is carried out in the way explained in Sec. 3. There is no iterative re-alignment performed here, although we have found that realignment brings further improvement to the system.

4.2.3. Testing

Testing is carried out over the target language, using the KIEL-test set. The two neural networks - for attribute estimation and for phoneme state estimation - are cascaded together. The decoding is performed using DNN-HMM system described in Sec. 3.

4.3. Baseline methods

The performance of attribute-based adaptation scheme is compared with a number of other popular schemes given below.

4.3.1. GMM-HMM

A GMM-HMM system is used for phoneme recognition task. The Kaldi TIMIT-s5 recipe is used for implementing mono-phone tri-state models. The input is MFCCs with delta and delta-delta coefficients. The system is trained entirely from the KIEL-adaptation set.

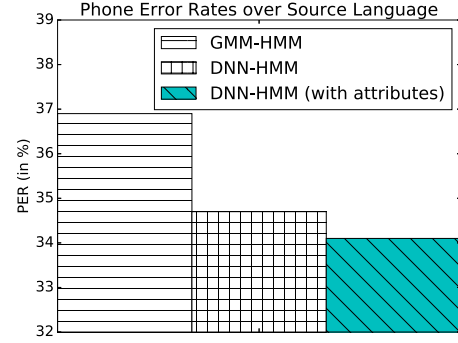


Fig. 2: Phoneme Error Rate (PER), in %ge, over the source language (English)

4.3.2. DNN-HMM

A conventional DNN-HMM is trained for phoneme recognition over the target language, using only the adaption set. It has four hidden layers consisting of 500, 500, 200 and 500 neurons. Notably, its architecture is the same as the cascaded DNN system in the proposed scheme, sans the layer of attributes. The input to the DNN and the decoding scheme is the same as that for the proposed method described in Secs. 2.1 and 3. This method forms the baseline of how well a conventional DNN-HMM system performs if trained over a limited target language data, when no cross-language knowledge is used.

4.3.3. DNN-HMM transfer with SHL without attributes

A popular approach to model transfer in DNNs is to transfer all the hidden layers and retrain the output layer for the target language [7]. The conventional DNN-HMM described above (Sec. 4.3.2) is trained for phoneme recognition over the source language, using the training set. For adapting the DNN to the target language, all but the output layer are transferred. The output layer is re-initialised randomly, and is trained using the adaptation data. The weights of the hidden layers are kept same. The weight update scheme is same as described in Sec. 3. This method conveys how effective the conventional scheme for cross-language transfer is without the use of attributes.

5. EXPERIMENTS

5.1. Phoneme Recognition over Source Language

In order to evaluate the capability of attributes for faithfully capturing the acoustic information, experiments have been performed for phoneme recognition over the source language. The only data available for training is the training dataset of the source language, i.e. the TIMIT-train set, and the test data

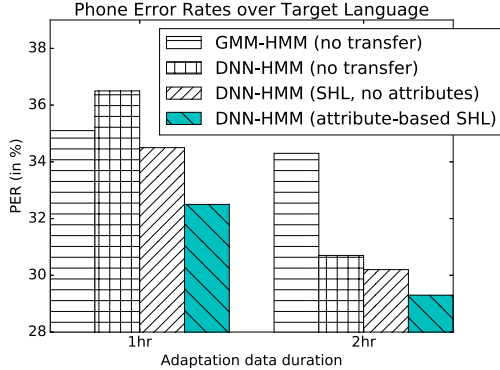


Fig. 3: Phoneme Error Rate (PER), in %ge, over the target language (German)

is the test dataset of the source language, i.e. the TIMIT-test set. The attribute extraction as well as the phoneme recogniser layers are trained using the TIMIT-train set. For training attribute estimation, all 61 phonemes of TIMIT are used; but for training and evaluation for phoneme recognition, they are reduced to 48 and 39 phonemes, respectively. The phoneme error rates obtained with different methods are presented in Fig. 2.

As expected, DNN-HMM system outperforms GMM-HMM system. The advantage of using attributes is visible even for the source language. The improvement here may not be significant but the main advantage here is that attribute-based SHL are meaningful, and yet perform on par with those without attributes. This, in itself, is a desirable property [24].

5.2. Phoneme Recognition over Target Language

For evaluating the efficacy of the proposed attribute-based cross-language model transfer method, experiments are performed for phoneme recognition over the target language. The training data is available in the form of the TIMIT-train and the KIEL-adaptation sets, and the KIEL-test set is used for testing. The phoneme error rates (PER) obtained using different methods are presented in Fig. 3. For very small adaptation data available, i.e. of 1 hour duration, the conventional DNN-HMM without model transfer performs worse than GMM-HMM system; but with larger adaptation data, the former surpasses the latter. DNN-HMM with SHL performs better than the one without model transfer. However, the proposed DNN-HMM with attribute-based SHL performs better than all other systems for both sizes of the adaptation set. The effect of changing the size of adaptation set is also visible. More adaptation data yields better performance in the form of reduction in PER. The improvement brought about by the proposed method is more significant in case of the smaller (i.e. 1hr) adaptation set. These results suggest the advantage of using meaningful attribute-based SHL. They can reliably transfer knowledge across different languages and

Table 2: F-values for attribute detection

	TIMIT-train	KIEL-adaptation	
		1hr	2hr
VOC	91.5	93.9	95.1
CONS	89.0	90.9	92.8
CONT	88.6	91.5	93.0
OBSTR	92.6	91.9	93.7
STR	90.4	87.9	89.8
VOICE	71.1	76.2	78.6
SON	97.2	97.1	97.6
STOP	80.2	74.6	78.3
LOW	86.1	95.7	96.5
HIGH	89.7	89.3	92.1
LAB	70.4	80.2	82.7
COR	89.1	86.4	88.0
DOR	75.1	85.6	87.5
RTR	86.1	81.2	82.9
NAS	84.3	92.4	94.0
LAT	75.2	53.7	63.0
RHO	67.4	33.7	53.4
RAD	64.2	60.0	64.7
SIL	93.3	96.2	97.1

can be retrained effectively on the target language. Hence, they lead to better performance even with a small adaptation data.

5.3. Attribute Recognition

It will be insightful to see how well the proposed system performs for the intermediate task of attribute extraction. Tab. 2 shows the F-values for recognising different attributes by the proposed DNN system. These values are measured frame-wise. A threshold level is used to decide if the feature is detected by the system, and is obtained from the training (or adaptation) data. The first column of F-values shows the performance of system described in Sec. 5.1 over the TIMIT-test set, and the subsequent ones show the performance of systems described in Sec. 5.2 over the KIEL-test set. We can observe that the relative performance over the features remains similar for all the three systems. This implies that some features are harder to detect than others. Also, the performance for all the features improves when we have more adaptation data, as seen from last two columns.

6. CONCLUSION

In this paper, we proposed using meaningful attributes for the hidden layers of DNN to be transferred across different domains. Phonological features are attributes shared by world's languages, and hence, provide a principled way to transfer models across languages. Experiments showed that DNN layers transferred with the help of attributes lead to effective and

reliable transfer, leading to better performance in phoneme recognition over the target language. The proposed method gave better results compared to other conventional methods including the one involving transfer of hidden layers without attributes. The improvements were seen even when very little data is available for adapting to the target language. In addition, the experiments also showed that attribute-based speech recognition provides interpretability to the hidden layers of DNN without compromising on the recognition accuracy.

In future, we would like to extend our work to word recognition and large vocabulary continuous speech recognition. We shall explore how well the phonological features work when phoneme level transcriptions are not available, since most large datasets are transcribed at word level only. The effect of using triphone HMMs and re-alignment remain to be explored as well. We would also like to investigate the attribute recognition accuracy in lines of zero shot recognition. This will include analysing how effectively the attributes are detected from unseen phonemes. Toward this end, we could even use models like convolutional neural networks to see how well can they discover invariant acoustic correlates associated with each attribute. It will be interesting to explore attribute-based transfer for DNNs in other areas of machine learning as well.

7. REFERENCES

- [1] Dong Yu and Li Deng, *Automatic Speech Recognition: A Deep Learning Approach*, Springer, 2014.
- [2] Tanja Schultz and Alex Waibel, “Language-independent and language-adaptive acoustic modeling for speech recognition,” *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, 2001.
- [3] Frantisek Grézl, Martin Karafiát, and Karel Veselý, “Adaptation of multilingual stacked bottle-neck neural network structure for new language,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 7654–7658.
- [4] Frantisek Grézl, Ekaterina Egorova, and Martin Karafiát, “Further investigation into multilingual training and adaptation of stacked bottle-neck neural network structure,” in *IEEE Workshop on Spoken Language Technology*, 2014, pp. 48–53.
- [5] Yu Zhang, Ekapol Chuangsuwanich, and James Glass, “Extracting deep neural network bottleneck features using low-rank matrix factorization,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 185–189.
- [6] Partha Lal and Simon King, “Cross-lingual automatic speech recognition using tandem features,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 12, pp. 2506–2515, 2013.
- [7] Jui Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7304–7308.
- [8] Peter Bell, Joris Driesen, and Steve Renals, “Cross-lingual adaptation with multi-task adaptive networks,” in *INTERSPEECH*, 2014, pp. 21–25.
- [9] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, “Unsupervised Cross-Lingual Knowledge Transfer in DNN-Based LVCSR,” in *IEEE Workshop on Spoken Language Technology*, 2012, pp. 246–251.
- [10] Dinesh Jayaraman and Kristen Grauman, “Zero-Shot Recognition with Unreliable Attributes,” in *Advances in Neural Information Processing Systems*, 2014, pp. 1–4.
- [11] Bernardino Romera-paredes and Philip H. S. Torr, “An embarrassingly simple approach to zero-shot learning,” in *Proceedings of the 32nd international conference on Machine learning (ICML)*, 2015, pp. 2152–2161.
- [12] Saumya Jetley, Bernardino Romera-Paredes, Sadeep Jayasumana, and Philip Torr, “Prototypical Priors : From Improving Classification to Zero-Shot Learning,” in *British Machine Vision Conference*, 2015.
- [13] Simon King and Paul Taylor, “Detection of phonological features in continuous speech using neural networks,” *Computer Speech and Language*, vol. 14, no. 4, pp. 333–353, 2000.
- [14] Patricia Scanlon, Daniel P. W. Ellis, and Richard B. Reilly, “Using broad phonetic group experts for improved speech recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 803–812, 2007.
- [15] Sabato Marco Siniscalchi, Dong Yu, Li Deng, and Chin Hui Lee, “Exploiting deep neural networks for detection-based speech recognition,” *Neurocomputing*, vol. 106, pp. 148–157, 2013.
- [16] Aren Jansen and Partha Niyogi, “Point process models for spotting keywords in continuous speech,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 8, pp. 1457–1470, 2009.
- [17] Milos Cernak, Blaise Potard, and Philip N Garner, “Phonological Vocoding using Artificial Neural Networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4844–4848.

- [18] Tasha Nagamine, Michael L. Seltzer, and Nima Mesgarani, “Exploring how deep neural networks form phonemic categories,” in *INTERSPEECH*, 2015, pp. 1912–1916.
- [19] O. Çetin, M. Magimai-Doss, K. Livescu, A. Kantor, S. King, C. Bartels, and J. Frankel, “Monolingual and crosslingual comparison of tandem features derived from articulatory and phone MLPS,” in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2007, pp. 36–41.
- [20] Sebastian Stuker, Florian Metze, Tanja Schultz, and Alex Waibel, “Integrating Multilingual Articulatory Features Into Speech Recognition,” in *EUROSPEECH*, 2003, pp. 1033–1036.
- [21] Sabato Marco Siniscalchi, Dau Cheng Lyu, Torbjørn Svendsen, and Chin Hui Lee, “Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 875–887, 2012.
- [22] Aditi Lahiri and Henning Reetz, “Distinctive features: Phonological underspecification in representation and processing,” *Journal of Phonetics*, vol. 38, no. 1, pp. 44–59, 2010.
- [23] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, “The kaldi speech recognition toolkit,” in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.
- [24] Shawn Tan, Khe Chai Sim, and Mark Gales, “Improving the interpretability of deep neural networks with stimulated learning,” in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 617–623.