Interspeech 2018
2-6 September 2018, Hyderabad

Phone Recognition using a Non-Linear Manifold with Broad Phone Class Dependent DNNs

Mengjie Qian, Linxue Bai, Peter Jančovič and Martin Russell

Department of Electronic, Electrical & Systems Engineering, The University of Birmingham, UK
{mxq486,lxb190,p.jancovic,m.j.russell}@bham.ac.uk

Abstract

Although it is generally accepted that different broad phone classes (BPCs) have different production mechanisms and are better described by different types of features, most automatic speech recognition (ASR) systems use the same features and decision criteria for all phones. Motivated by this observation, this paper proposes a two-level DNN structure, referred to as a BPC-DNN, inspired by the notion of a topological manifold. In the first level, several small separate BPC-dependent DNNs are applied to different broad phonetic classes, and in the second level the outputs of these DNNs are fused to obtain senone-dependent posterior probabilities, which can be used for frame level classification or integrated into Viterbi decoding for phone recognition. In a previous paper using this approach we reported improved frame classification accuracy on the TIMIT corpus compared with a conventional DNN. The contribution of the present paper is to demonstrate that this advantage extends to full phone recognition. Our most recent results show that the BPC-DNN achieves reductions in error rate relative to a conventional DNN of 16% and 8% for frame classification and phone recognition, respectively.

Index Terms: manifold learning, phone classification, speech recognition, neural network, broad phone classes

1. Introduction

State-of-the-art automatic speech recognition (ASR) systems use a single deep neural network (DNN) to define a mapping f from the "acoustic space" A to the space P of vectors of phone (or senone) posterior probabilities, which are integrated into the Viterbi decoder for speech recognition. Although this is a single continuous mapping, in practice the DNN is trained to approximate a discontinuous function f whose outputs typically jump between 0 and 1 across phone state boundaries. Therefore, it may be advantageous to think of f as a *set* of continuous functions $\{f_1, \cdots, f_J\}$, with each function f_j defined on a subset $A_j \subset A$ of the acoustic space and $\bigcup_{j=1}^{J} A_j = A$. In this case the appropriate mathematical structure is a non-linear topological manifold. There have been few studies that have shown the benefits of coding the acoustic speech signal in a non-linear manifold space [1, 2, 3].

In context of speech analysis, the manifold structure provides a tool to exploit the fact that different phonetic classes employ different production mechanisms and are best described by different types of features. Intuitively, one might hope that the subsets A_j correspond to broad phonetic categories. The idea of phone-dependent feature extraction is well-established. For example, while vocal tract resonance frequencies provide a natural description of vowels, unvoiced consonants are better described in terms of duration and mean energies in key frequency bands [4, 5, 6, 7, 8, 9, 10]. There are also a number of studies that use BPC-dependent classifiers to focus on subtle differences between phones within a BPC [11].

A two-level *linear* computational model motivated by these considerations is presented in [12]. The first level comprises a set of discriminative linear transforms, one for each of a set of overlapping broad phone classes (BPCs), that are used for feature extraction. The transforms are obtained using variants of linear discriminant analysis (LDA). Each transform is applied to an acoustic feature vector and k-nearest neighbour methods are used to estimate probabilities of BPCs and phones, which are then combined in the second level to estimate posterior probabilities and hence to classify the acoustic vector. This two-level linear classifier obtained slightly better results compared to a single transform on frame-level phone classification experiments on TIMIT [13].

Inspired by these observations, in our previous study [14] we introduced a two-level *non-linear* model, referred to as a BPC-DNN. Our premise was that it would be advantageous to replace a single 'global' DNN with several BPC-dependent DNNs. In the first level of the BPC-DNN, several small, separate DNNs were applied to different BPCs. For each BPC, a DNN was trained to map acoustic features onto a vector of posterior probabilities of the phones or senones within the BPC, plus an "outside-BPC" class. In the second level the outputs of these DNNs were passed as input to another DNN, the fusion network, which transformed them into a single phone or senone posterior probability vector, which was used for frame level classification. The BPC-DNN is related to Wu and Gales' multi-basis adaptive neural network (MBANN) [15], in which parallel component DNNs correspond to different speaker types.

An obstacle to the application of a topological manifold model to acoustic speech analysis is the need to cover the acoustic space A with phonetically meaningful subsets A_i on which the "feature extraction" transforms f_i are defined. In the approach described here this problem is avoided by applying the BPC-dependent DNN for a particular BPC to the whole of A but only mapping frames corresponding to phones in the BPC to the correct phone class. All other frames are mapped to the "outside-BPC" category.

It was shown in [14] that the BPC-DNN model gives statistically significant improvements in phone-classification of feature vectors, compared with a single global DNN. The contribution of the present paper is to extend this work to full phone recognition by passing the output of the fusion network to a Viterbi decoder.

2. Broad Phone Classes

Broad phonetic classes are defined in terms of a common articulatory strategy that is used in their production. In a BPC-DNN the component DNNs correspond to BPCs or combinations of BPCs. The elements of the TIMIT 49 phone set are partitioned

10.21437/Interspeech.2018-1376

into 8 non-overlapping BPCs, referred to as $\{G_1, \cdots, G_8\}$ in Table 1. Consonants are divided into "plosives", "strong fricatives", "weak fricatives" and "nasals/flaps" $(G_1, \cdots, G_4)$, liquids and glides are considered as "semi-vowels" (G_5), the vowels are grouped into "short vowels" and "long vowels" (G_6, G_7).We also define 6 'super' phone classes $\{G_9, \cdots G_{14}\}$, which are the union of two or more BPCs from $\{G_1, \cdots, G_8\}$, to combine broad classes that are frequently confused. These are the BPCs from [12].

Table 1: *Broad phone classes and super classes.*

Group	Phonetic class	Phone labels
G_1	Plosive	/b/, /d/, /g/, /k/, /p/, /t/
G_2	Strong fricative	/ch/, /jh/, /s/, /sh/, /z/, /zh/
G_3	Weak fricative	/dh/, /f/, /hh/, /th/, /v/
G_4	Nasal/Flap	/dx/, /en/, /m/, /n/, /ng/
G_5	Semi-vowel	/el/, /l/, /r/, /w/, /y/
G_6	Short vowel	/aa/, /ae/, /ah/, /ax/, /eh/, /ih/, /ix/, /uh/
G_7	Long vowel	/ao/, /aw/, /ay/, /er/, /ey/, /iy/, /ow/, /oy/, /uw/
G_8	Silence	/cl/, /epi/, /q/, /sil/, /vcl/
G_9	$G_5 \cup G_6 \cup G_7$	Semi-vowel, Short vowel, Long vowel
G_{10}	$G_1 \cup G_3$	Plosive, Weak fricative
G_{11}	$G_5 \cup G_6$	Semi-vowel, Short vowel
G_{12}	$G_5 \cup G_7$	Semi-vowel, Long vowel
G_{13}	$G_6 \cup G_7$	Short vowel, Long vowel
G_{14}	$G_1 \cup G_2 \cup ... \cup G_8$	All phones

3. A Two-Level Broad Phone Class DNN (BPC-DNN)

3.1. Upper level: BPC-dependent DNNs

The input to the i^{th} DNN in the upper-level of the BPC-DNN is a filter-bank feature vector in context, and the output is a set of $n_i + 1$ posterior probabilites, one for each of the n_i phone or senone classes in the i^{th} BPC plus an additional "not in the BPC" probability. In this way all of the training data is used to train each upper-level DNN and the need to identify a subset A_i of the acoustic feature space A corresponding the i^{th} BPC is avoided. We explored the use of different combination of BPCs.

3.2. Lower level: single fusion Network

The outputs of all of the upper-level BPC-dependent DNNs are concatenated to form the input to the lower-level fusion network. The output nodes of the fusion DNN correspond to the posterior probabilities of all of the phones or senones in the complete phone set. In the present implementation the fusion network has a single hidden layer. The structure of a two-level BPC-DNN is shown in Figure 1.

4. Experiments with phone-level alignments

This section compares phone-level frame classification and phone recognition results obtained with a conventional single global DNN and a two-layer BPC-DNN system described in Section (3).

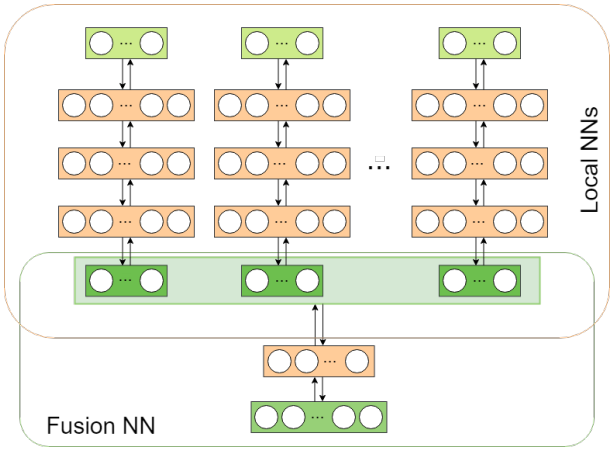

Figure 1: *The structure of a BPC-dependent neural network.*

4.1. Data and features

Experiments were performed on the 16kHz TIMIT speech corpus [13] which has 6300 sentences recorded from 630 speakers. The SA recordings were excluded, hence there are 3696 utterances and 192 utterances in the training and test set, respectively. The 61 phone set used in TIMIT labels is mapped to the 49 phone set [16] for training and testing then be further reduced to 40 [16] for evaluating the results.

4.2. Baseline systems

The baseline ASR model (BASE_MONO1) is a hybrid deep neural network - hidden Markov model (DNN-HMM) trained using the Kaldi toolkit [17]. The speech was encoded as MFCC vectors plus delta and delta-delta coefficients (39 parameters) and used to train a *single state* monophone HMM system (hence there is no distinction between phone and senone labels). The alignments from this monophone model were subsequently used to train a baseline DNN with three hidden layers, each with 1024 nodes.

The inputs to the DNN were 26 dimensional filter-bank features with a context of 11 frames (i.e. $\pm$5 frames). The output layer is a softmax layer with 49 nodes corresponding to the posterior probabilities of each of the 49 phones.

We evaluated this model in terms of both frame accuracy and percentage phone recognition error. A bi-gram language model trained on the transcriptions in the training set was used for phone recognition. The results are shown in Table 4 (first row).

4.3. BPC-DNN systems (one state per phone)

Each BPC-DNN corresponds to a set of BPCs, which determine the number of BPC-dependent DNNs in the upper layer. We considered five different sets, referred to as $D_1, \cdots, D_5$ in Table 2. D_1 consists of the 8 non-overlapping BPCs from Table (1), while D_2 to D_5 also includes some of the "super groups".

The input features for each local network are the same 26 dimensional filter-bank features with a context of $\pm$5 frames. Each of the local BPC-dependent DNNs in the upper layer has 3 hidden layers each with 256 nodes. The phone alignment from the baseline model is modified to train the BPC-dependent DNNs in the upper layer. For BPC G_i, the labels in the alignment which correspond to phones in G_i are kept unchanged

Table 2: *Sets $D_1, \cdots, D_5$ of BPCs used to train BPC-DNNs.*

Broad phone class	Experimental setup				
	D_1	D_2	D_3	D_4	D_5
G_1 - G_8	X	X	X	X	
G_9		X	X		
G_{10}			X	X	X
G_{11}				X	X
G_{12}				X	X
G_{13}				X	X
G_{14}					X
# of local DNNs	8	9	10	12	13
total out-nodes in mono1	57	80	92	116	165
total out-nodes in mono3	155	222	256	324	471

while the labels which are not in G_i are all mapped to the same "out-of-group" phone label.

For example, D_1 consists of the 8 BPCs $G_1, \cdots, G_8$ containing $6, 6, 5, 5, 5, 8, 9$ and 5 phones, respectively. Since each BPC-dependent DNN in the upper layer also contains a "not in this class" output, the total number of outputs from the upper layer (inputs to the lower layer) is $49 + 8 = 57$. The number of output nodes from all of the BPC-dependent DNNs in the upper layer for each D_i are shown in Table 2. The original monophone alignment was used to train the fusion network and thus there are 49 output nodes in this network. We only use one hidden layer in the fusion network, but we explored the influence of different number of hidden nodes. We also explored the use of context in the posterior probability vectors that are input to the fusion network.

The results of experiments using 32 and 64 hidden nodes in the fusion DNN, without context (32_0 and 64_0) and with a context of ± 5 phone posterior probability vectors in the input layer (32_5 and 64_5) are shown in Figure 2. The horizontal red dashed line shows the results obtained using the baseline global DNN. All of the results are with respect to the standard 40 phone TIMIT set. The figures show that the two-layer DNNs, with BPC-dependent neural networks in the upper layer and a fusion DNN in the lower layer, outperform the baseline global DNN both in terms of frame accuracy and phone error rate.

Table 3: *Number of parameters (millions) in the 1-state and 3-state per phone-HMM BPC-HMM systems. The corresponding single DNN baseline systems have 2.44 and 2.54 parameters, respectively.*

# of states per phone	Fusion NN	D1	D2	D3	D4	D5
1 state	32_0	1.66	1.87	2.08	2.50	2.72
	32_5	1.68	1.90	2.11	2.53	2.76
	64_0	1.66	1.87	2.08	2.50	2.72
	64_5	1.70	1.93	2.14	2.58	2.83
3 states	32_0	1.69	1.91	2.13	2.56	2.81
	32_5	1.74	1.99	2.21	2.66	2.96
	64_0	1.70	1.93	2.14	2.57	2.83
	64_5	1.80	2.07	2.31	2.78	3.13

For frame classification (Figure 2, top figure), the two-layer BPC-DNN systems corresponding to D_4 and D_5 with a context of ± 5 frames in the input to the fusion DNN achieve a reduction in frame classification error rate of approximately 13% relative to the baseline system. The only cases where the performance of the two-layer system is poorer than the single global baseline network are the experiments for D_1 without context in the input to the fusion layer (these are the blue and green columns on the left of Figure 2). However, these two-layer BPC-DNNs have fewer parameters than the baseline DNN.

For phone recognition (Figure 2, bottom figure), the phone error rates for the two-layer BPC-DNNs are again lower than for the baseline system except for those systems corresponding to D_1 without context in the input to the fusion network. The best system corresponds to D_5, with 32 units in the hidden layer of the fusion DNN and a context of ± 5 frames for the input to the fusion DNN. This system achieves a reduction in error rate of approximately 4% relative to the baseline.

The number of parameters for each system are shown in the top part of Table 3.

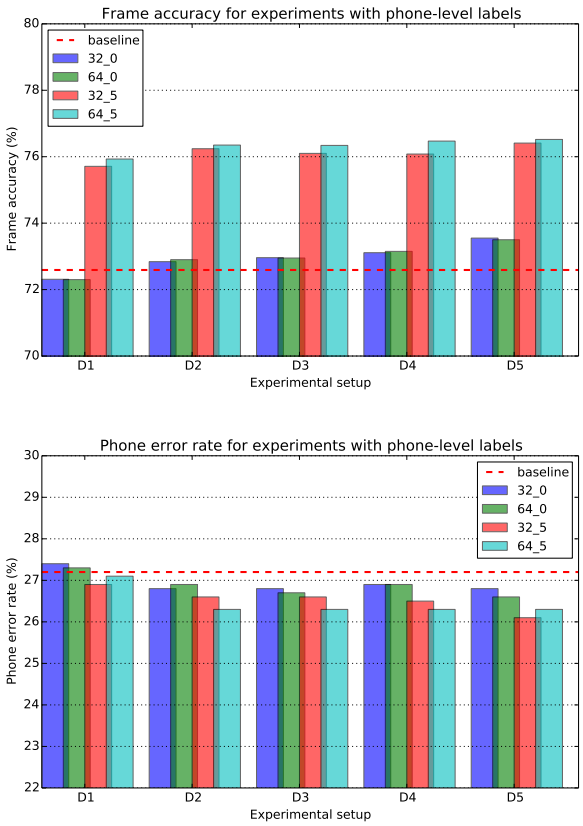

Figure 2: *Percentage Frame accuracy (top) and percentage phone error rate (bottom) for experiments with 32 and 64 hidden nodes and a context of ± 0 and ± 5 frames in the input to the fusion DNN, using the phone-level labels.*

5. Experiments with state-level alignments

For phone recognition, in contrast with the systems described in Section (4), it is normal to use 3 states per phone-level HMM.

This section presents results for frame classification and phone recognition using senone-level alignments obtained with 3 state phone-level HMMs.

5.1. Baseline

A baseline DNN (BASE_MONO3) was trained using the same features and hidden layer structures as in section 4.2, but with different alignments. We trained a monophone GMM-HMM system with 3 states per phone-HMM. In this system there are 147 phone states and the output layer of the BASE_MONO3 DNN has 147 nodes representing the posterior probabilities of these 147 HMM states. The results are shown in Table 4 (second row).

5.2. BPC-DNN systems (three states per phone)

The alignments from sub-section (5.1) that were used to train the baseline were also used in training the BPC-DNNs. When training a local BPC-specfic DNN, the HMM states in the alignment corresponding to phones that are not in this group were again mapped to a "out-of-group" node. We explored the use of combinations BPC-dependent DNNs for the different combinations of BPCs ($D_1, \cdots, D_5$) from Table 2. For each D_i, the total number of output nodes of the local networks are shown in the last row of Table 2. Again we only used one hidden layer in the fusion network, but with different numbers of hidden nodes (32 or 64), and contexts of 0 or ± 5 frames on the input layer of the fusion network.

The frame accuracies and the phone error rate are shown in Figure 3. The number of parameters in each system are shown in the bottom part of Table 3.

Table 4: *Percentage frame accuracies (%FAC) for 147, 49 and 40 targets, and percentage phone error rates (%PER) for the baseline DNN models (base_1 and base_3) and the best performing BPC-DNNs (BPC_1 and BPC_3) with 1 and 3 states per phone.*

	%FAC-147	%FAC-49	%FAC-40	%PER
base_1	-	69.69	72.59	27.2
base_3	61.69	69.64	72.49	26.7
BPC_1	-	73.98	76.52	26.1
BPC_3	66.35	74.46	77.00	25.1

For frame classification (Figure 3, top figure), all of the two-layer BPC-DNN systems outperform the baseline. The best performance, corresponding to D_5 with a context of ± 5 frames in the input to the fusion DNN, achieves a reduction in frame classification error rate of approximately 16% relative to the baseline system. However this BPC-DNN system also has approximately 23% more parameters than the baseline. For BPC-DNN $D4$ with no context the number of parameters is similar to the baseline and the reduction in frame classification error is approximately 4%.

For phone recognition (Figure 3, bottom figure), all of the two-layer BPC-DNN systems again outperform the baseline. With a context of ± 5 the BPC-HMM systems corresponding to D_2, D_3, D_4 and D_5 all achieve a reduction in phone error rate of approximately 6% relative to the baseline. In the cases of D_2 and D_3 this is achieved with fewer parameters than the baseline.

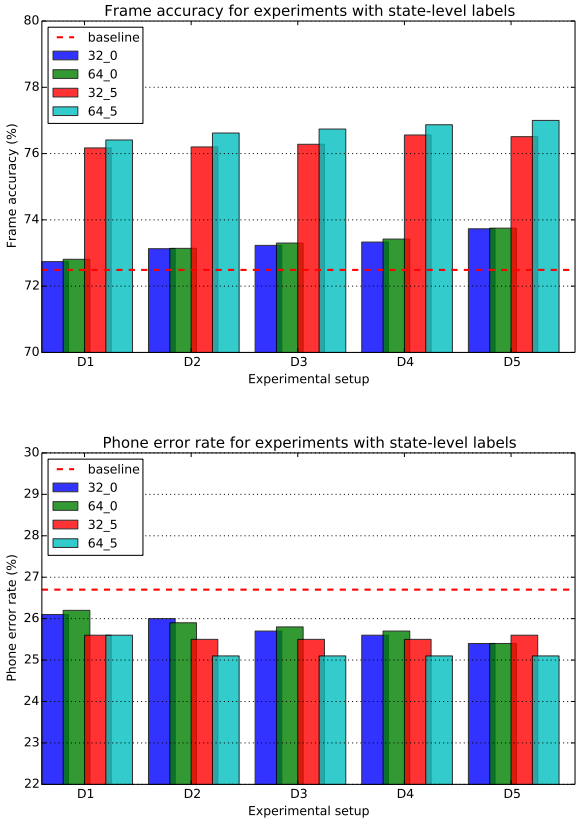

Figure 3: *Percentage Frame accuracy (top) and percentage phone error rate (bottom) for experiments with 32 and 64 hidden nodes and a context of ± 0 and ± 5 frames in the input to the fusion DNN, using the state-level labels.*

6. Conclusions and discussion

This paper describes ongoing research into the application of DNN-based models inspired by the notion of topologial manifold to speech analysis and recognition (BPC-DNNs). Our premise is that such a model is of interest because it reflects the fact that different types of speech sound, corresponding to different modes of production, lend themselves naturally to different types of acoustic analysis. The main conclusion from this work is that the improvement in frame phone classification accuracy previously reported using BPC-DNNs can be extended to phone recognition. Specifically, we obtain a reduction in phone error rate of 6% relative to a conventional DNN using a BPC-DNN with fewer parameters.

The BPC-DNN only approximates a topological manifold structure because the "local" mappings f_i are implemented by DNNs defined on the whole acoustic space A. This raises the question of whether better performance, and more insight, could be obtained with a more faithful manifold structure in which A is covered by proper subsets A_i. For example, if $A_i \cap A_j \neq \emptyset$ and $v \in A_i \cap A_j$ then $f_i(v)$ and $f_j(v)$ could be interpreted as alternative analyses of v from the perspective of different BPCs, and therefore potentially different production mechanisms.

7. References

[1] A. Jansen and P. Niyogi, "Intrinsic fourier analysis on the manifold of speech sounds," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.

[2] A. Subramanya and J. A. Bilmes, "Entropic graph regularization in non-parametric semi-supervised classification," in *Advances in Neural Information Processing Systems*, 2009, pp. 1803–1811.

[3] Y. Liu and K. Kirchhoff, "Graph-based semi-supervised learning for phone and segment classification." in *INTERSPEECH*, 2013, pp. 1840–1843.

[4] F. Li, A. Menon, and J. Allen, "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," *The Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2599–2610, 2010.

[5] F. Li, A. Trevino, A. Menon, and J. B. Allen, "A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise," *The Journal of the Acoustical Society of America*, vol. 132, no. 4, pp. 2663–2675, 2012.

[6] K. Stevens and S. Blumstein, "Invariant cues for place of articulation in stop consonants," *The Journal of the Acoustical Society of America*, vol. 64, no. 5, pp. 1358–1368, 1978.

[7] J. M. Heinz and K. N. Stevens, "On the properties of voiceless fricative consonants," *The Journal of the Acoustical Society of America*, vol. 33, no. 5, pp. 589–596, 1961.

[8] L. Raphael, "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English," *The Journal of the Acoustical Society of America*, vol. 51, no. 4B, pp. 1296–1303, 1972.

[9] L. Wilde, "Analysis and synthesis of fricative consonants," Ph.D. dissertation, Massachusetts Institute of Technology, 1995.

[10] A. Khasanova, J. Cole, and M. Hasegawa-Johnson, "Detecting articulatory compensation in acoustic data through linear regression modeling," in *Proc. Interspeech,* Singapore, 2014.

[11] P. Scanlon, D. Ellis, and R. 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.

[12] H. Huang, Y. Liu, L. ten Bosch, B. Cranena, and L. Boves, "Locally learning heterogeneous manifolds for phonetic classification," *Computer Speech and Language*, vol. 38, pp. 28–45, 2016.

[13] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Univ. Pennsylvania, Philadelphia, PA, 1993.

[14] L. Bai, P. Jancovic, M. Russell, P. Weber, and S. Houghton, "Phone classification using a non-linear manifold with broad phone class dependent dnns," *Proc. Interspeech 2017*, pp. 319–323, 2017.

[15] C. Wu and M. J. F. Gales, "Multi-basis adaptive neural network for rapid adaptation in speech recognition," in *icassp15*, 2015.

[16] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, "The kaldi speech recognition toolkit," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.