INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Transfer Learning with Bottleneck Feature Networks for Whispered Speech Recognition

Boon Pang Lim[1], *Faith Wong*[2], *Yuyao Li*[2], *Jia Wei Bay*[2]

[1]Institute for Infocomm Research, Singapore
[2]Nanyang Girls' High School, Singapore
bplim@i2r.a-star.edu.sg[1]

Abstract

Previous work on whispered speech recognition has shown that acoustic models (AM) trained on whispered speech can somewhat classify unwhispered (neutral) speech sounds, but not vice versa. In fact, AMs trained purely on neutral speech completely fail to recognize whispered speech. Meanwhile, recipes used to train neutral AMs will work just as well for whispered speech, but such methods require a large volume of transcribed whispered speech which is expensive to gather. In this work, we propose and investigate the use of bottleneck feature networks to normalize differences between whispered and neutral speech modes. Our extensive experiments show that this type of speech variability can be effectively normalized. We also show that it is possible to transfer this knowledge from two source languages with whispered speech (Mandarin and English), to a new target language (Malay) without whispered speech. Furthermore, we report a substantial reduction in word error rate for cross-mode speech recognition, effectively demonstrate that it is possible to train acoustic models capable of classifying both types of speech without needing any additional whispered speech.
Index Terms: speech recognition, whispered speech, deep neural networks, bottleneck feature networks, multilingual bottlenecks, low-resource speech recognition.

1. Introduction

Whispering, though a common mode of speech communication, is not often studied. In recent times, speech recognition has taken off commercially, driving various applications such as mobile voice search [1]. Even though the speech modality offers advantages in streamlining access to ever more ubiquitious computers, its use can be limited in certain social situations where it would be inconvenient to speak out loud – there is a need for machines to recognize whispered speech in such settings. Currently, speech recognition systems rely on many hundreds or thousands of hours of training data to achieve acceptable performance. Furthermore, training algorithms are still not robust enough to deal with variabilities in speech and signal. This is especially so for whispered speech – it has been shown that while reasonably accurate AMs can be trained from whispered speech to recognize the same type of speech, models trained from non-whispered (*neutral*) speech cannot recognize whispered speech. Given the lack of sizable whispered speech corpora, this makes constructing whispered speech recognizers particularly difficult or expensive.

In this work we present our findings on using bottleneck feature networks to transfer knowledge of speaking modes from one language to another. Our method trains bottleneck feature (BNF) networks by mapping corresponding whispered and neutral sounds for the same triphone to the same target senone label. We show that by doing so, the BNF network effectively maps similar speech sounds of otherwise different speech modes onto the same feature space – this mapping can be used to train a whispered speech recognizer for a new target language without needing additional whispered speech.

2. Related Work

In [2], a comparison of speech recognizers respectively trained on whispered speech and neutral speech – as tested with both types of speech – show a severe degradation in speech recognition performance when recognizing whispered speech with a neutral speech model. This degradation is not as pronounced if an AM trained from whispered speech were to be used to recognize neutral speech. This result has been consistent for experiments done in several languages for which there is a sizeable corpus of parallel whispered speech, such as Serbian [3], Japanese [4], Mandarin [5] and English [6, 7]. The pattern is also consistent for different types of models, both Gaussian Mixture Models (GMM) trained with generative or discriminative methods, and even for Deep Neural Network (DNN) based AMs. At the same time other experiments with large vocabulary speech recognition demonstrate that it is feasible to build speech recognizers for whispered speech using a conventional training recipe. However such methods would require a large and potentially expensive database of transcribed whispered speech.

Many attempts have been made to compensate for the differences between neutral and whispered speech in order to improve whispered speech recognition. In [6], several different methods for adapting a neutral AM to whispered speech were tried, including variants of MAP, MLLR and eigenvoices. In [8], the authors present both feature and model-based methods – feature based methods include the application of Shift Frequency [9] features and Vocal Tract Length Normalization (VTLN), using reduced bandwidth filterbanks during feature extraction. They also propose a model-based approach of using the Vector Taylor Series (VTS) algorithm to generate pseudo-whispered speech from neutral speech to facilitate AM training. Most of these approaches demonstrate improvement in adapting neutral AMs to whisper. Their later work in [10] proposed to use Denoising Autoencoders (DAE) to directly map whispered speech frames to neutral speech frames and vice versa. The authors used different DAEs for different broad phone classes, and tried to map both single frames and short segments of different speech modes to each other. These techniques fail to show much further improvement beyond the earlier VTS approach. Further-

http://dx.doi.org/10.21437/Interspeech.2016-250

more, all of these approaches have yet to fully compensate for the differences between whispered and neutral speech.

Perhaps the difficulties encountered with methods for speech mode neutralization can be accounted for by the non-linear differences between whisper and neutral speech. Such differences have been studied in languages such as Serbian [11], Japanese [12], English [13], Mandarin [5] and many other languages. Among these works there is consistent agreement that whispered speech differs from neutral speech in three major ways – in terms of the lack of voicing, in terms of the reduced spectral tilt, and also in terms of lengthening of vowels. Such differences not only span across multiple speech frames, but are possibly non-linear in the feature space, and cannot be captured by simple affine transformations at that level. This suggests that it may perhaps be necessary to use more computationally powerful functions to map neutral to whispered speech or vice versa. Thus, this motivates our use of DNNs, specifically deep bottleneck networks to learn this mapping.

Bottleneck feature networks and their variants are well-studied in the literature. They evolved out of earlier work on Tandem features [14], in which a shallower neural network is used to estimate phone posterior probabilities in order to provide complementary information to raw features. These estimated posterior vectors are concatenated with traditional speech recognition features such as MFCC and used in tandem to build AMs. Later, [15] introduced the use of bottleneck feature networks that contain a bottleneck layer with a limited number of hidden neurons. Their networks accepted TRAPS-based features [16] as input and used phoneme targets in the output layer, and were designed to perform dimensionality reduction on raw features, while retaining sufficient information to accurately classify phone posteriors.

Recently, the success of Deep Neural Networks in acoustic modeling [17] have led to so-called deep bottleneck features. In [18], DNN based bottleneck features were trained by simply introducing a bottleneck into a specific layer of a DNN. They investigated three potential sources of improvements to deep BNF networks – through pretraining, increasing the number of hidden layers, and increasing the number of classification targets used during training. The authors showed that pretraining using Restricted Boltzmann Machines (RBN) improved results for shallower networks and also made the improvements more stable. They also showed improvements by going from monophone to clustered triphone targets. As an alternative to pretraining, the layer-wise training using autoencoders was proposed in [19]. Use of low-rank matrix factorization to create bottleneck from an already trained network was proposed in [20]. Further refinements and variants to the basic bottleneck feature setup include stacking – in which multiple bottleneck networks are concatenated to further improve signal fidelity in the feature stream [21, 22]. Some of these works also leverage data from other language corpora in order to train speech recognizers in low resource conditions, effectively performing a form of multilingual transfer learning.

3. Bottleneck Feature Training

Our approach to DNN-HMM hybrid ASR with bottlenecks is illustrated in Figure 1 – it is essentially a large network separated into a feature compensation frontend and an acoustic model backend. Raw speech features for a single frames are first spliced to provide additional context information. These are fed into a neural network with a bottleneck layer at the output. The output of this layer are BNFs – they are again spliced

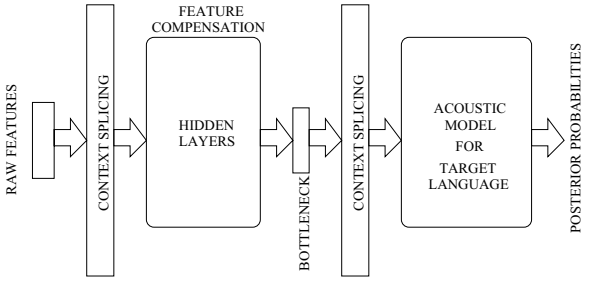

Figure 1: Hybrid DNN-HMM Speech Recognition with Bottleneck Features for Speech-mode Normalization.

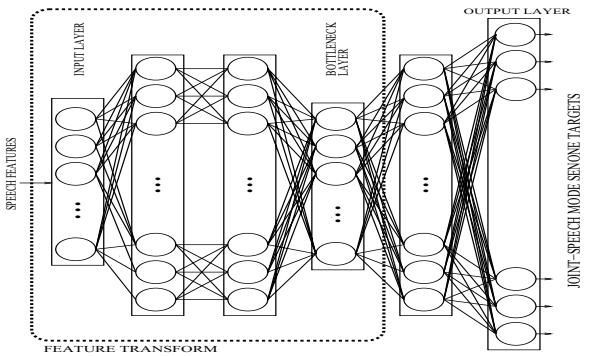

Figure 2: Use of Joint Speech Mode Senone Targets for Bottleneck Feature Training.

to generate an even longer temporal context. Spliced BNF features are used to train the backend acoustic model.

Figure 2 illustrates the procedure for training bottleneck feature networks. During training we use a neural network with multiple hidden layers and a narrow bottleneck layer that divides it into two sections. The neurons in the final output layer correspond to senone targets for a joint speech mode model. First, a joint-mode GMM acoustic model is trained from labelled whispered and neutral speech. As the joint-mode AMs are bootstrapped triphone clustering decision trees are built and corresponding forced alignments of the training data are generated. These are used together to train the bottleneck network. The bottleneck can be thought of to restrict information flow through the network. Maintaining a high classification accuracy at the final output layer retains information pertinent to speech sound classification at the bottleneck layer. Even though the network is not explicitly trained to differentiate between neutral and whispered speech, with careful tuning, information relevant to speech mode is lost through the bottleneck. This effectively projects speech of different modes onto similar regions in the BNF space, allowing a speech recognizer trained from features in the projected space to recognize either speech mode. Since senone labels are used at the output layer during BNF training, frame classification accuracy and cross entropy loss here give us a good gauge of how salient the BN features are for speech recognition, albeit for a mismatched language.

4. Experiments and Discussion

4.1. Corpora and Acoustic Models

We used two parallel whispered speech corpora and one neutral speech corpus in our experiments. The English whispered

Corpus	Language	Hrs	#C. Triphones
iWhisper	Mandarin	32.5	3720
wTIMIT	English	52.1	5326
MASS	Malay	52.2	11372

Table 1: Corpora and Trained Acoustic Models

Model-Type	GMM-MLE	DNN
Neut. WER	20.6	12.0
Whis. WER	70.5	71.1

Table 2: Malay (Neutral) Acoustic Model ASR performance.

TIMIT corpus contains 48 speakers speaking and whispering all 450 phonetically balanced TIMIT sentences. The Mandarin iWhisper-Mandarin corpus contains 80 speakers both reading and whispering distinct sets of 100 sentence collected from online web resources. The Malay language MASS corpus was used for training the large vocabulary speech recognizers [23]. In addition, we collected a small set of whispered Malay speech in order to validate our systems. We recruited 4 male and 3 female speakers and asked them to each read and whisper utterances from a list of 50 sentences. These sentences were selected at random from the MASS corpus. Recordings were made in a quiet environment, using an Audio-Technica ATH150COM USB headset with a 16kHz sampling rate with 16-bits per sample. This gave us slightly over 15 minutes of test data in each speech mode.

These corpora were used to train two joint-mode acoustic models in English and Mandarin, and one neutral Malay acoustic model. We used 13-dimension MFCCs in conjunction with a 1-dimension autocorrelation-based pitch feature [24], and applied up to 3rd order deltas to get 56 dimension features. The feature extraction used 10ms framesteps with a 25ms Hamming window. The joint-mode acoustic models are trained using both types of labelled speech together. A standard recipe from Kaldi [25] was used. We bootstrapped the AM from monophones, then proceeded to clustered triphones, applying LDA and MLLT to train a GMM model using Maximum Likelihood Estimation (MLE). At each stage the best model trained according to a cross-validation set was used to generate better alignments. The acoustic model for Malay was trained in a similar fashion. Statistics on these corpora and the clustered triphone set size for the correspondingly best trained acoustic model are summarized in Table 1. Cross-mode speech recognition experiments for these models were previously reported in [5] and thus not included here.

The WERs for the neutral Malay AM tested on held out whispered and neutral Malay speech is shown in Table 2. Unsurprisingly, the best performing model for matched speech modes is DNN-based. Although the GMM model is slightly better for mismatched speech modes, the error rate is high enough so that the relative difference in performance is quite small. Thus, we choose to use DNN models for most of our remaining experiments; this gives us baseline WERs of 71.1% (testing on whisper) and 12.0% (testing on neutral) to serve as a comparison with later experiments. As a comparison, WER of 14.1% was previously reported in [23] for this corpus, albeit for a different test set. The ideal speech-mode compensation method should be able to bring down the WER to as close that for neutral speech without sacrificing the fidelity for the original speech.

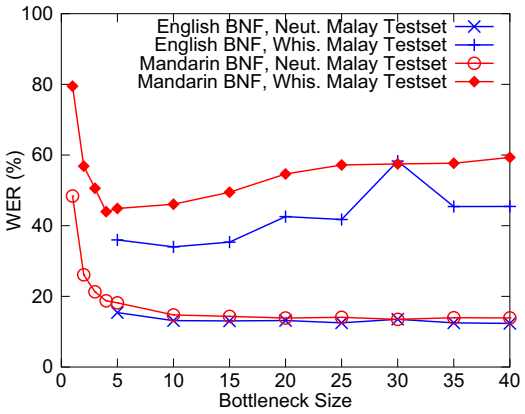

Figure 3: ASR performance for Neutral Malay AM: Effect of different BN Sizes.

4.2. Effect of Bottleneck Dimension

This set of experiments examines the effect of having different bottleneck sizes. The previously trained joint-mode AMs are used to generate forced alignments for a small subset of the training data. This was controlled to be around 15 hours of speech in total for each language to speed up the training process. The alignments are used to train BNF networks.

We used identical features as those used in the joint-mode AM training, but did not apply deltas so as to more directly control context effects. These 14-dimension features were spliced with the adjacent 5 left and 5 right frames to generate 154 dimension features that span 125 ms temporally. The resulting raw bottleneck features are spliced again using the adjacent 5 left and 5 right BNF frames and fed to the target language acoustic model for training. This finally results in a temporal context of 225 ms for DNN training. In our pilot experiments we did not find any significant differences using pre-training so this step was omitted. All BNF networks had three hidden layers with 1024 neurons in the layers before the bottleneck, and one hidden layer after the bottleneck. They were trained using Stochastic Gradient Descent to convergence, using learning rates from between 0.0002 and 0.0008 and applying halving after improvements start to fall below a threshold. Figure 3 compare the accuracy of the systems with bottlenecks trained either solely from English or from Mandarin, when it is applied to the Malay language. Cross-entropies and frame accuracies for selected bottleneck sizes is shown in Table 3. We observe that the BNF network trained from parallel whispered English speech with a bottleneck size of 30 has unusually high WERs on whispered Malay. This correlates with a higher cross entropy and lower frame accuracy in the bottleneck layer. We suspect that too small a learning rate was chosen for this particular setup, leading to the network not converging well even after 20 iterations.

The results show that larger bottleneck widths benefit speech recognition in general, leading to improved recognition accuracies for neutral speech. However this also results in letting too much information pertaining to the speech mode through, hurting whispered speech recognition. A good balance is obtained with low-dimension bottlenecks, but the exact point differs depending on language setup. For the English-based BNFs this seems to be around 5 to 10 nodes; for Mandarin this seems to be around 5 nodes. A detailed look at frame accura-

BN Size	5	25	30	35
BNF Fr Acc.	44.98	53.23	47.80	55.01
BNF Xent	2.492	1.854	2.220	1.746
DNN Fr Acc.	32.22	44.68	45.09	45.77
DNN Xent	3.264	2.421	2.408	2.370
Ne Acc.	15.41	12.51	13.52	12.51
Wh Acc.	35.98	41.76	58.28	45.42

Table 3: Cross-Entropy, Frame Accuracy and Malay AM performance for different English BN Sizes.

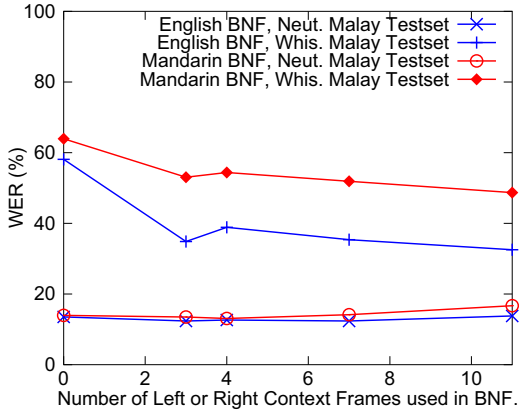

Figure 4: Effect of Context Splicing (245ms Total Context)

cies and cross entropies evaluated over the cross-validation set shows that there is a good correlation between high frame accuracy and low cross entropy at the backend, with final system accuracy. However good metrics at the BNF does not necessarily translate to the best performance.

4.3. Effect of Context Splicing

Context-splicing allows us to integrate information from adjacent frames so as to improve DNN classification accuracy. In these experiments we investigate the effect of using different amounts of context in the frontend and backend while keeping the total amount of temporal information integrated constant. We dropped the pitch feature in these experiments and a bottleneck size of 25 was used. Figure 4 plots the WER for systems built using different context splicing options. The x-axis labels the number of left or right contexts used in the frontend BNF network. A value of 5 indicates a total of 11 frames spliced or a temporal context of 125ms. In all systems the number of contexts used in the frontend BNF and the backend DNN sum to 11, thus they integrate up to 245ms of temporal context as an input to the backend AM.

The plots indicate that for most part the effect of context splicing is minimal when testing with neutral speech, but it has a great effect when testing with whispered speech. The best results on neutral speech are obtained when the amount of context used at each stage is roughly the same. Whispered speech recognition keeps improving as greater amounts of temporal context is integrated at the BNF. It is interesting to note that even though the same amount of context is presented to the backend AM, there can be some signal degradation for neutral speech recognition. Perhaps it is necessary to increase the bottleneck sizes in these cases to allow more information useful to classification to be let through.

BNF	Type of Acoustic Model					
	GMM			DNN		
	Neut.	Whis		Neut.	Whis	
	WER	WER	Rel.	WER	WER	Rel.
none	20.5	79.5	-	12.0	71.1	-
En	23.6	44.4	44.2	11.9	36.0	49.4
Ma	20.8	63.4	20.2	13.2	49.5	30.4
En+Ma	21.6	46.1	42.1	12.0	41.8	41.2

Table 4: Improved Recognition of Whispered Malay speech with BNFs.

4.4. Training Whispered Speech Recognizers without Whispered Speech

Finally, we present our findings on combining corpora from other languages in order to train whispered speech recognizers for a new target language. In this set of experiments, 13-MFCC and 1-pitch feature was used as raw features and third-order delta coefficients were extracted to get 56 dimension features. Adjacent frames were spliced (4 left, 4 right) to obtain a 616-dimension vectored inputs used to train bottleneck feature networks. A multilingual BNF setup similar to [26] was used to combine data from both languages for training. Malay acoustic models were trained for each BNF setup by filtering all the neutral Malay speech training data, then using those features to first bootstrap a GMM model (trained using maximum likelihood) and then a DNN model (using stochastic gradient descent). The final systems were tested on both neutral and whispered Malay speech. Table 4 summarizes our results. The performance of our neutral Malay speech acoustic model is replicated in the first line for reference, and the columns marked `Rel.` contain the relative WER reduction on whispered test speech. Our results demonstrate that the BNF networks can drastically improve whispered Malay speech recognition. Furthermore using any of the trained BNF feature extractors with different types of acoustic modelling show between 20% to 49% relative word error rate reduction on whispered speech, but without increasing the error rate for neutral speech, suggesting that the BNF networks are compensating for differences in whispered and neutral speech. In this sense the compensatory differences between neutral and whispered speech is transferred from a speech corpus in a different language to a new one. Currently, combining data from multiple languages does not seem to help with system performance. We plan to further investigate this anomaly.

5. Conclusion

We have presented a set of experiments to demonstrate the feasibility of applying BNF networks to recognizing whispered speech. This approach shows promise – it is possible to train whispered speech recognizers for languages which do not have any existing labeled whispered speech corpora. We plan to further investigate different language options and regularization methods for network training in future work.

6. Acknowledgements

The authors would like to thank Dr Sunil Sivadass, Dr Nancy Chen and Ms Gu Yuling for helpful discussions and feedback on the manuscript and experiments.

7. References

[1] I. McGraw, R. Prabhavalkar, R. Alvarez, M. G. Arenas, K. Rao, D. Rybach, O. Alsharif, H. Sak, A. Gruenstein, F. Beaufays, and C. Parada, "Personalized speech recognition on mobile devices," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[2] T. Itoh, K. Takeda, and F. Itakura, "Acoustic analysis and recognition of whispered speech," in *Acoustics Speech and Signal Processing*, vol. 1, 2002, pp. 389–392.

[3] D. T. Grozdic, B. Markovic, J. Galic, and S. T. Jovocic, "Application of neural networks in whispered speech recognition," *Telfor Journal*, vol. 5, pp. 103–105, 2013.

[4] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, pp. 139–152, 2005.

[5] P. X. Lee, D. Wee, B. P. Lim, N. F.-Y. Chen, and B. Ma, "A whispered Mandarin corpus for speech technology applications," in *INTERSPEECH*, 2014.

[6] B. P. Lim, "Computational differences between whispered and non-whispered speech," Ph.D. dissertation, University of Illinois at Urbana Champaign, Dec 2010.

[7] S. Ghaffarzadegan, H. Boril, and J. Hansen, "UT-VOCAL effort II: Analysis and constrained-lexicon recognition of whispered speech," in *in Proc. IEEE ICASSP*, 2014, pp. 2544–2548.

[8] S. Ghaffarzadegan, H. Boril, and J. H. L. Hansen, "Model and feature based compensation for whispered speech recognition," in *INTERSPEECH*, 2014.

[9] H. Boril, "Robust speech recognition: Analysis and equalization of Lombard effect in czech corpora," Ph.D. dissertation, Czech Technical University in Prague, Czech Republic, 2008.

[10] S. Ghaffarzadegan, H. Boril, and J. Hansen, "Generative modeling of pseudo-target domain adaptation samples for whispered speech recognition," in *ICASSP*, 2015.

[11] S. T. Jovicic and Z. Saric, "Acoustic analysis of consonants in whispered speech," *Journal of Voice*, vol. 22, no. 3, pp. 263–274, 2008.

[12] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, pp. 139–152, October 2003.

[13] R. W. Morris, "Enhancement and recognition of whispered speech," Ph.D. dissertation, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, August 2003.

[14] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *ICASSP*, 2000.

[15] F. Grézl, M. Karafiát, S. Kontár, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *ICASSP*, vol. 4. IEEE, 2007, pp. IV–757.

[16] H. Hermansky and S. Sharma, "Temporal patterns (traps) in asr of noisy speech," in *in Proc. ICASSP*, 1999, pp. 289–292.

[17] G. Hinton, L. Deng, D. Yu, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. S. G. Dahl, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.

[18] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *INTERSPEECH*. International Speech Communication Association, August 2011.

[19] J. Gehring, Y. Miao, F. Metze, and A. Waibel, "Extracting deep bottleneck features using stacked auto-encoders," in *ICASSP*, May 2013, pp. 3377–3381.

[20] Y. Zhang, E. Chuangsuwanich, and J. Glass, "Extracting deep neural network bottleneck features using low-rank matrix factorization," in *ICASSP*, 2014.

[21] ——, "Language ID-based training of multilingual stacked bottleneck features," in *INTERSPEECH*, 2014.

[22] F. Grezl, M. Karafiat, and K. Vesely, "Adaptation of multilingual stacked bottleneck neural network structure for new languages," in *ICASSP*, 2014.

[23] T. P. Tan, H. Li, E. K. Tang, X. Xiong, and E. S. Chng, "MASS: A Malay language lvsr corpus resource," in *in Proc. of 2009 Oriental COCOSDA Int. Conf.*, 2009, pp. 26–30.

[24] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. Klein and K. K. Palival, Eds. Elsevier, 1995.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vasely, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[26] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *ICASSP*, 2013, pp. 7304–7308.