Speech Recognition with Temporal Neural Networks

Payton Lin[1], *Dau-Cheng Lyu*[2], *Yun-Fan Chang*[1], *Yu Tsao*[1]

[1] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[2]ASUS Headquarters, Advanced Technology Division, Kauhsiung, Taiwan

{paytonlin,she2113,yu.tsao}@citi.sinica.edu.tw, Daucheng_Lyu@asus.com

Abstract

Raw temporal features were derived from extracted temporal envelope bank (referred to as "Tbank"). Tbank features were used with deep neural networks (DNNs) to greatly increase the amount of detailed information about the past to be carried forward to help in the interpretation of the future.

Index Terms: raw temporal features, temporal neural network

1. Introduction

For speech feature learning and for speech recognition, the goal has been condensed to the use of primitive spectral or possibly waveform features [1]. Deep neural networks (DNNs) exploit information in neighboring frames and by modeling tied context-dependent (CD) states [2]. The whole network is discriminatively fine-tuned to predict the target hidden Markov model (HMM) states. Back-propagation repeatedly adjusts the weights of the connections in the network so as to minimize a measure of the difference between the actual output vector of the net and the desired output vector [3]. Gradient descent can be used for fine-tuning the weights, but this works well only if the initial weights are close to a good solution [4]. "Raw" spectral features not only retain more information (including possibly redundant or irrelevant ones), but also enable the use of convolution and pooling operations [1]. Convolutional neural networks (CNNs) have showed success in achieving translation invariance for many image processing tasks [5]. It is suitable for image processing because the same image pattern can appear at any position in an image [6]. Research for acoustic modeling using deep belief networks (DBNs) are currently exploring alternative input representations that allow DNNs to **see** more of the relevant information in the sound-wave, such as very precise coincidences of onset times in different frequency bands [7].

The treatment of the time dimension of speech by DNN-HMM and Gaussian mixture model (GMM)-HMMs alike is viewed as a very crude way of dealing with the intricate properties of speech [8]. Enhancing the features causes the network to be less robust to mismatched conditions, e.g. SNR or channel variations, because it **sees** fewer variations in the data during training [9]. DNNs have the ability to generate internal representations that are robust [10] with respect to variability **seen** in the training data. Overfitting can be reduced by using "dropout" to prevent complex co-adaptations on the training data [11]. These improvements on DNN architecture and learning were necessary to push the features back to the raw level of acoustic measurements [1]. Yet, it is believed that features better than mel-scale filter-bank may be discovered in the near future to further boost CD-DNN-HMMs [12].

The present study expands on previous research to explore alternative input representations that allow DNNs to **hear** more of the relevant information in the sound-wave, such as very precise coincidences of onset times in different temporal bands. High levels of speech recognition have been achieved with a new sound processing strategy for multielectrode cochlear implants [13]. No specific features of the speech input such as fundamental or formant frequencies of voiced sounds, were extracted explicitly. Speech pattern recognition can be achieved with primarily temporal cues [14]. These cues are often expressed over diverse time spans that would benefit from different lengths of analysis widows in speech analysis and feature extraction [15]. The speech information related to the fundamental frequency of speech, on the other hand, can be encoded in electric repetition rate as in the speech processor [16]. Further, the discriminative cues among separate speech classes are often distributed over a reasonably long temporal span, which often crosses neighboring speech units [15].

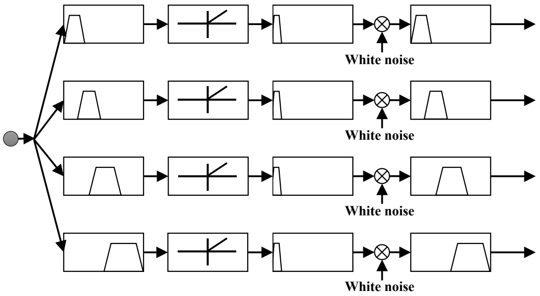

Figure 1: *Temporal envelope extraction with 4 bands.*

"Raw" temporal features were derived from extracted temporal envelope bank (Tbank) to confront the problem of dimensionality reduction: finding meaningful low-dimensional structures hidden in their high-dimensional observations [17]. Temporal envelopes of speech were extracted from broad frequency bands and were used to modulate noises of the same bandwidths [14], as shown in Figure 1. The CD-DNN-HMM has the ability to generate more invariant and selective features at higher hidden layers [12]. Hidden layers can be considered as increasingly complex feature transformations and the final softmax layer as a log-linear classifier making use of the most abstract features computed in the hidden layers [18]. For comparison with raw spectral features: 40-dimensional log mel filter-bank (referred to as "Fbank"), Tbank features were derived with 40 coefficients distributed on a mel-scale (Fig. 2).

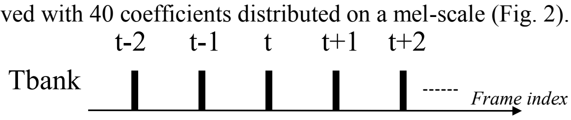

Figure 2: *Time-locked static feature parameters of Tbank.*

The unique characteristic of speech lies primarily in its temporal dimension- in particular, in the huge variability of speech associated with the elasticity of this temporal dimension [15]. The time functions are expanded by orthogonal polynomial representations [19]. First- and second-order derivative features were used in an investigation of DNN for noise robust speech recognition [9]. While these dynamic features are useful [12], difficulty was found for learning or improving delta-like features [1]. Figure 3 evaluates the effect of appending dynamic temporal derivatives to Tbank+Δ+ΔΔ.

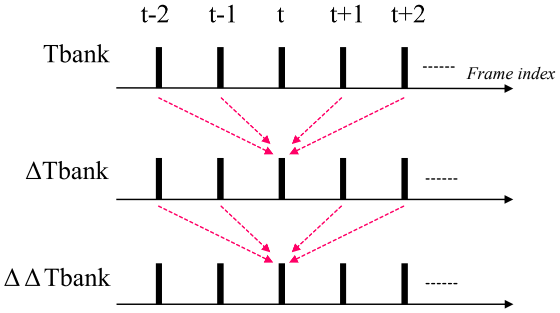

Figure 3: *Combinations of the static Tbank+Δ+ΔΔ features.*

2. Experiments

To evaluate the speech recognition performance of the DNN-HMM, a series of experiments was performed on Aurora-4 [20], a medium vocabulary task based on the Wall Street Journal (WSJ0) corpus. The experiments were performed with the 16 kHz clean training set consisting of 7137 utterances from 83 speakers. The evaluation set was Test Set 1 (clean data), derived from WSJ0 5k-word closed vocabulary test set which consists of 330 utterances from 8 speakers.

The baseline GMM-HMM system consisted of context-dependent HMMs with 2032 senones and 16 Gaussians per state trained using maximum likelihood estimation. The input features were 13-dimensional mel frequency cepstral coefficient (MFCC) features and cepstral mean subtraction was performed. The 13-dimensional MFCC features were spliced in time taking a context size of 7 frames, followed by de-correlation and dimensionality reduction to 40 using Linear Discriminant Analysis (LDA) [21]. The resulting features were further de-correlated using maximum likelihood linear transform (MLLT) [22]. These models were also used to align the training data to create senone labels for training the DNN-HMM system. Decoding was performed with the WSJ0 trigram language model.

DNNs were trained using either raw Fbank or Tbank features. Utterance-level mean and variance normalization was performed. The input layer was formed from a context window of 11 frames creating an input layer of 440 visible units for the network. DNNs had 5 hidden layers with 2048 hidden units in each layer and the final soft-max output layer had 2032 units, corresponding to the senones of the HMM-system. The networks were initialized using layer-by-layer generative pre-training and then discriminatively trained using twenty-five iterations of back propagation. A learning rate of 0.16 was used for the first 15 epochs and 0.004 for the remaining 10 epochs, with a momentum of 0.9. Back propagation was done using stochastic gradient descent in mini batches of 512 training examples. DNN dropout training was not used in preliminary experiments because dropout essentially reduces the capacity of the DNN and thus can improve the generalization of the resulting model [9].

3. Temporal neural network

Deep learning, representation learning, and unsupervised feature learning sets an important goal of automatic discovery of powerful features from raw input data independent of application domain [1]. Because the temporal structure of the HMM is maintained [10], a temporal neural network (TNN) refers to when temporal envelope alignment [23] is used during the steps [8] of converting and denoting, prior to fine-tuning the DNN and re-estimating the transition probabilities.

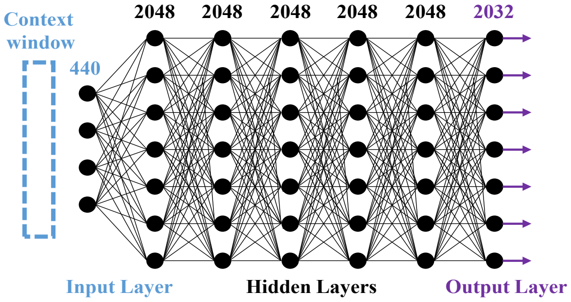

Figure 4: *Structure of temporal neural network (TNN).*

There is a vast space to explore in the deep learning framework using insights gained from temporal-centric generative modeling [8]. A learning algorithm creates internal representations that are demonstrably the most efficient way of using the preexisting connectivity structure [24]. The present study evaluates the learning algorithm that is capable of learning the underlying constraints that characterize a temporal domain simply by being shown examples from the temporal domain. DBNs are probabilistic generative models with multiple layers of stochastic hidden units above a single bottom layer of observed variables that represent a data vector [8]. The generative pre-training creates many layers of feature detectors that become progressively more complex [7]. The generative model learned during pre-training helps overfitting, even when using models with very high capacity and can aid in the subsequent optimization of the recognition weighs [8].

New improvements on DNN architectures and learning are needed to push the features even further back to the raw level of acoustic measurements [1]. Due to automatic learning of representations [1], temporal envelope extraction should pull the features back even further to the rawest level of acoustic measurements [ex: 13, 14, 16]. A previous correspondence has investigated the automatic recognition of cochlear implant-like spectrally reduced speech, which is synthesized from subband temporal envelopes of the original clean speech [25]. For comparison with raw temporal features (Fig. 1), the envelope could also be used to modulate a sine-wave generated at the center of the analysis band, as shown in Figure 5.

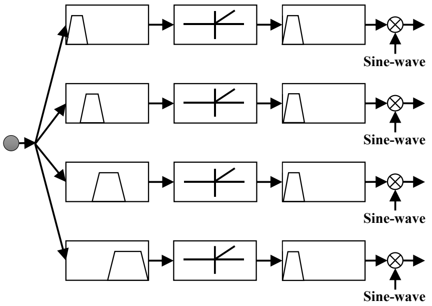

Figure 5: *Sine-wave synthesis with 4 bands.*

4. Effects of the number of bands

Figure 6 displays performance for temporal features as a function of the number of temporal bands. Triphone HMMs were trained on the clean training set. The clean test data was split into 1, 2, 4, 6, 8, 16, or 24 bands for temporal envelope extraction (Fig. 1) or sine-wave synthesis (Fig. 5). Sidebands of sine-waves provided a periodic temporal structure that enhanced discrimination capabilities of acoustic units for 6 - 8 bands. However, the overlapping spectral sidebands generated from amplitude modulation caused performance to decrease from 80.78 to 70.52% when the number of bands increased from 8 to 24. For temporal envelopes, the sidebands of white-noise carriers were masked by carrier spectra, limiting the delicate structure of triphones and speech attributes with 6 and 8 bands. However, the contiguous bands of white-noise enabled performance for temporal envelopes to increase monotonically as the number of bands increased from 1 to 24, with accuracy improving from 5.08 to 87.22%. Findings were similar to other databases [23], with higher accuracy attributed to using a balanced database [20] or performing LDA+MLLT instead of appending first- and second- order time derivatives.

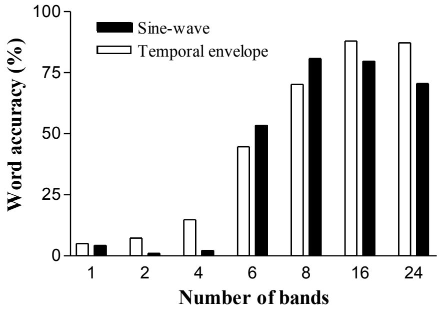

Figure 6: *GMM-HMM performance on Tbank features.*

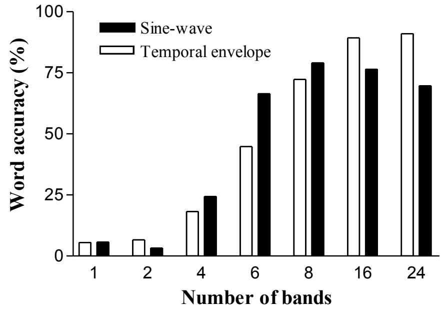

Figure 7: *CD-DNN-HMM performance on Tbank features.*

Figure 7 displays DNN performance for temporal features when trained using raw Fbank spectral features. For sine-wave synthesis, word accuracy decreased from 79.02 to 69.70% as the number of bands increased from 8 to 24. Performance for sine-wave synthesis increased up to 24.38% with 4 bands. CD-DNN-HMM improved performance compared to the GMM-HMM system in Fig. 6, where performance was limited to 2.22% with 4 bands. Performance for temporal envelopes increased monotonically from 1 to 24 bands, with accuracy improving from 5.53 to 91.05%. CD-DNN-HMMs provide flexibility of using arbitrary features [12]. These internal representations become increasingly insensitive to small perturbations in the input with increasing network depth [10].

5. Effects of temporal dynamics

The presentation of a dynamic temporal pattern in only a few broad spectral regions was sufficient for the recognition of speech [14]. High speech recognition performance could be achieved with only three time-varying bands of noise representing the complex spectral patterns of speech. Although the previous section showed temporal waveform envelope provides significant information for DNN speech recognition, nearly perfect speech recognition could not be computed under these parameters of greatly reduced spectral information.

Figure 8 shows appending dynamic features improved accuracy for temporal envelopes from 18.21% (Fig. 7) to 22.29% with 4 bands. This technique broadened the DNN input layer from 440 to 1320 visible units. Figure 9 shows decreasing the number of coefficients in Tbank from 40 to 24 improved accuracy for sine-waves from 30.30% (Fig. 7) to 34.95% with 4 bands. This techniques reduced the DNN input layer from 1320 to 792 visible units. Although incorporating temporal derivatives improved Tbank feature robustness, the presentation of a dynamic pattern in only a few broad spectral regions was still not sufficient for the recognition of speech.

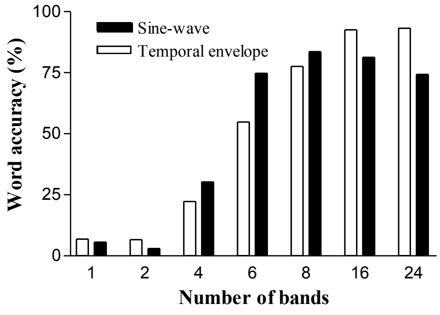

Figure 8: *CD-DNN-HMM performance on Tbank+Δ+ΔΔ.*

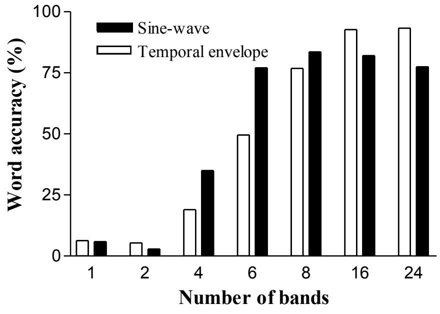

Figure 9: *CD-DNN-HMM on 24-dimensional Tbank+Δ+ΔΔ.*

6. Learning temporal representations

Performance is improved by incorporating information about the environment into DNN training [9]. This ability allows us to just feed in heterogeneous data collected under different environments and expect DNNs to reduce the mismatch and be robust to the variation [12]. A simplified representation of speech is able to support relatively high levels of open-set recognition [13]. Spectral information was removed from speech by replacement of the frequency region with a band-limited noise [14]. In this respect, performing automatic speech recognition with HMMs trained on spectrally reduced speech could assess the relevance of spectrally reduced speech as a model for speech recognition [25]. Table 1 compares word error rate (WER%) with HMMs trained on various Tbank features. Because the temporal structure of the HMM is maintained [10], Table 1 shows temporal waveform envelope provides significant information for TNN speech recognition in the deep learning framework.

Tbank Features	WER% (GMM)	WER% (TNN)
8 band envelopes	6.99	4.99
8 sine-waves	5.47	3.59
24 band envelopes	5.03	3.47
24 sine-waves	7.55	4.22

Table 1: *Performance (WER%) with matched Tbank training.*

Table 2 compares WER% at different steps [8] between a CD-DNN-HMM baseline and a TNN. Performance for GMM-HMMs improved when trained and tested on 24 band envelopes compared to baseline MFCC features. These Tbank models were then used to align the training data to create senone labels for training a TNN. Next, matched Tbank training with a TNN reduced WER% on the Tbank evaluation set compared to the Fbank training set results with CD-DNN-HMMs. Finally, Table 2 shows the TNN gives fewer errors if trained and tested on the same Fbank features as the CD-DNN-HMM baseline. Using a better alignment to generate training labels for the DNN can improve the accuracy [8].

System/Features	WER%
GMM-HMM on MFCC	5.08
GMM-HMM on Tbank	5.03
CD-DNN-HMM on Tbank	8.95
TNN on Tbank	3.47
CD-DNN-HMM on Fbank	2.88
TNN on Fbank	2.63

Table 2: *Comparison of systems at different steps of training.*

By using mixed-bandwidth training data, the DNN learns to consider the differences in the wideband and narrowband input features as irrelevant variations [12]. To avoid a “hand-crafted” transformation of speech spectrogram, and the known loss of information from the raw speech data [1], Table 3 only combines Tbank features in steps [8] to generate a state-level alignment on the training set to preserve the frequency axis for other techniques [ex: 1, 5, 6]. Any improvements in modeling units that are incorporated into the CD-GMM-HMM baseline system, such as cross-word triphone models, will be accessible to the DNN through the use of the shared training labels [8].

Alignment Features	WER% (GMM)	WER% (DNN)
MFCC	5.08	2.88
+3 band envelopes	4.76	2.76

Table 3: *Temporal representation during state-level alignment*

Table 3 shows combining Tbank feature representation during steps to generate a state-level alignment allows the DNN to give fewer errors if trained and tested on raw Fbank features.

7. Discussion

Speech presents a difficult pattern recognition problem for the auditory system [14]. The human brain confronts the same problem in everyday perception, extracting from its high-dimensional sensory inputs-30,000 auditory nerve fibers or 10^6 optic nerve fibers-a manageably small number of perceptually relevant features [17]. The recognition of speech has been thought to require frequency-specific (spectral) cues [14]. There is nothing in the ear corresponding to the crystalline lens of the eye, and this not accidently, so to speak, but by the very nature of the case [26]. Nearly perfect speech recognition was observed under conditions of greatly reduced spectral information [14]. These novel finding imply that the process of speech perception makes use of time-varying acoustic properties that are more abstract than the spectra and speech cues typically studied in speech research [27]. The processing of acoustic structure at the level of the syllable is best described as rhythm detection, and speech rhythm is determined principally by the acoustic structure of amplitude modulation at relatively low rates in the signal [28]. Rather, the ultralow-frequency modulation envelopes in the order of 3 to 8 Hz are critical cues to intelligibility [29]. Deciphering the information from amplitude-modulated (AM) sounds is a well-understood process, requiring a phase locking of primary auditory afferents to the modulation envelopes [30]. Rhythm in speech is a property of the slow AM of the waveform, corresponding roughly to the AM associated with syllables [28]. These findings lend support to recent theories of speech encoding that state, contrary to conventionally thinking, that a detailed auditory analysis of the short-term acoustic spectrum is not essential to the speech code [29].

The present results complement existing studies of simplified acoustic representations of speech for normal hearing listeners and provide further insight into the minimal cues necessary for speech understanding [13]. Many defining features of speech sounds are rapid temporal transitions with durations well within the reversal window [29]. Research for acoustic modeling using DBNs are currently exploring ways of using recurrent neural networks (RNNs) to greatly increase the amount of detailed information about the past that can be carried forward to help in the interpretation of the future [7]. Although the amplitude spectrum of a waveform is unaffected by time reversal, the temporal envelopes, as well as the fine structure of the running spectrum, are highly distorted for such sounds [29]. It is evident therefore that it is useless to look for anything corresponding to the crystalline lens of the eye, and that our power of telling the origin of a sound must be explained in some different way [26].

8. Conclusions

We present an alternative input representation that allows deep neural networks to hear more of the relevant information in the sound-wave, such as very precise coincidences of onset times in different temporal bands. Band envelopes pulled the features back even further to the rawest level of acoustic measurements. New improvements on TNN architectures and learning are needed to push the features even further back, setting up a new goal[1] for deep learning.

9. Acknowledgements

This research was supported by National Science Council of Taiwan under contracts NSC101-2221-E-001-020-MY3.

[1] [Online] Available: http://angelsim.tigerspeech.com

10. References

[1] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M.L Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero, “Recent advances in deep learning for speech research at Microsoft*,” in Proc. ICASSP*, 2013, pp. 8604-8608.

[2] G.E. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, “Deep Neural Networks for acoustic modeling in speech recognition,” *IEEE Signal Process. Mag.*, 82, pp. 82-97, 2012.

[3] G.E. Hinton and R.R Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, 313 (5786), pp. 504-507, 2006.

[4] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, “Learning representations by back-propagating errors,” *Nature*, 323 (9), pp. 533-536, 1986.

[5] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, “Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition,” *in Proc. ICASSP*, 2012, pp. 4277-4280.

[6] O. Abdel-Hamid, L. Deng, and D. Yu, “Exploring convolutional neural network structures and optimization techniques for speech recognition,” *in Proc. Interspeech*, 2013, pp. 3366-3370.

[7] A. Mohamed, G.E. Dahl, and G.E Hinton, "Acoustic modeling using deep belief networks." *IEEE Trans. Audio Speech Lang. Process.*, 20 (1), pp. 14-22, 2012.

[8] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” *IEEE Trans. Audio Speech Lang. Process*., 20 (1), pp. 30–42, 2012.

[9] M.L. Seltzer, D. Yu, and Y. Wang, “An investigation of deep neural networks for noise robust speech recognition,” *in Proc. ICASSP*, 2013, pp. 7398-7402.

[10] D. Yu, M.L. Seltzer, J. Li, J.-T. Huang, and F. Seide, “Feature learning in deep neural networks-studies on speech recognition tasks,” *in Proc. ICLR*, 2013, pp. 1-9.

[11] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov, “Improving neural networks by preventing co-adaption of feature detectors,” *CoRR*, abs/1207.0580, 2012.

[12] J. Li, D. Yu, J.T. Huang, and Y. Gong, “Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM.” *IEEE SLT*, 2012, pp. 131-136.

[13] B.S. Wilson, C.C. Finley, D.T. Lawson, R.D Wolford, D.K. Eddington, and W.M. Rabinowitz, “Better speech recognition with cochlear implants,” *Nature*, 352, pp. 236–238, 1991.

[14] R.V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, “Speech recognition with primarily temporal cues,” *Science*, 270 (5234), pp. 303-304, 1995.

[15] L. Deng, and X. Li, “Machine learning paradigms for speech recognition: an overview,” *IEEE Trans. Audio, Speech & Lang.*, 21 (5), pp. 1060-1089, 2013.

[16] Y.C. Tong, R.C. Dowell, P.J. Blamey, and G.M. Clark, “Two-component hearing sensation produced by two-electrode stimulation in the cochlea of a deaf patient,” *Science*, 219, (4587), pp. 993-994, 1983.

[17] J.B Tenenbaum, V. de Silva, and J.C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, 290 (5500), pp. 2319-2323, 2000.

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

[19] S. Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Trans. Acoust. Speech Signal Process.*, 29 (2), pp. 254-272, 1981.

[20] N. Parihar and J. Picone, “Aurora working group: Dsr front end lvcsr evaluation au/384/02,” *in Institute for Signal and Information Processing Report*, 2002.

[21] K. Fukunaga, “Introduction to statistical pattern recognition,” *Academic Press*, 1972.

[22] M.J.F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *in Tech. Report*, CUED/F-INFENG/TR291, 1997, Cambridge University

[23] P. Lin, S.-S. Wang, and Y. Tsao, “Temporal information in tone recognition,” *IEEE ICCE*, Taiwan, 2015.

[24] D.H. Ackley, G.E. Hinton, and T.J. Sejnowski, T. J, “A learning algorithm for boltzmann machines,” *Cognitive Science*, 9 (1), pp. 147-169, 1985.

[25] C.-T. Do, D. Pastor, and A. Gaolic, “On recognition of cochlear implant-like spectrally reduced speech with MFCC and HMM-based ASR,” *IEEE Trans. Audio, Speech, and Lang. Process.*, 18 (5), pp. 1065-1068, 2010.

[26] L. Rayleigh, “Our perception of the direction of a source of sound,” *Nature*, 14, pp. 32-33, 1876.

[27] R.E. Remez, P.E. Rubin, D.B. Pisoni, and T.D. Carrell, “Speech perception without traditional speech cues,” *Science*, 212, pp. 947-950, 1981.

[28] U. Goswami, J. Thomson, U. Richardson, R. Stainthorp, D. Hughes, S. Rosen, and S.K. Scott, “Amplitude envelope onsets and developmental dyslexia: A new hypothesis,” *in Proc. Nat. Acad. Sci. USA (PNAS)*, 99 (16), pp. 10911-10916, 2002.

[29] K. Saberi, and D.R. Perrott, “Cognitive restoration of reversed speech,” *Nature*, 398, pp. 760, 1999.

[30] K. Saberi, and E.R. Hafter, “A common neural code for frequency- and amplitude-modulated sounds,” *Nature*, 374, pp. 537-539, 1995.