Temporal Alignment for Deep 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

***Abstract*—Alternative features were derived from extracted temporal envelope bank (TBANK). These simplified temporal representations were investigated in alignment procedures to generate frame-level training labels for deep neural networks (DNNs). TBANK features improved temporal alignments both for supervised training and for context dependent tree building.**

Keywords—alignment, temporal features, deep neural networks

I. INTRODUCTION

Time alignment presents the greatest problem for neural network (NN) based systems, since connectionist learning procedures are typically defined in terms of static pattern classification tasks [1]. The output units in the linear output layer represent the targets of classification [2]. A neural network is trained using a state alignment that provides a label for each frame in the training set [3]. They use hidden Markov model (HMM) technology to find the optimal time alignment based upon the output of the connectionist component of the system [4]. The most current implementation of this approach is based on the deep neural network (DNN), or neural networks with many layers using back-propagation learning [2]. In the embedded Viterbi training procedure, the training label of the samples are determined using forced alignment [5]. Using a better alignment to generate training labels for the DNN can improve the accuracy [6]. It was also confirmed that the lower the error rate of the system used during forced alignment to generate frame-level training labels for the neural net, the lower the error rate of the final neural-net-based system [7]. Furthermore, the poorly matched alignment has a negative impact on the resulting system accuracy [8].

For the highly non-convex optimization problem of DNN learning, it is obvious that better parameter initialization techniques will lead to better models since optimization starts from these initial models [2]. Without translation invariance, a neural net requires precise segmentation to align the input pattern properly [9]. The insertion and deletion errors are due to imperfect segmentation, and the substitution errors are due to imperfect classification [10]. The learning procedure should not require precise temporal alignment of the labels that are to be learned [9]. Target label data are always available in direct or indirect forms for such supervised learning [2]. The targets are obtained by using a baseline gaussian mixture model (GMM)-HMM system to produce a forced alignment [7]. These DNN models thus have implicit dependency on the original GMM used, as well as the features and context window size of that model [3]. Tuning the probabilities again only marginally improves the performance [6]. Therefore, the plan is to investigate alternative features, and the impact of iteratively realigning with larger, state-of-the art models from flat start context independent alignments [3].

Alternative features [11] were derived[1] from extracted temporal envelope bank (TBANK) to determine whether HMM models generated from band envelope [12] could be used to align the training data to create labels for training the DNN-HMM. Intuitively, if the labels are generated from a more accurate model the trained DNN should perform better [5]. Previously, raw TBANK features were constructed with a 25-ms Hamming window shifted with 10-ms steps to evaluate the structure of the amplitude envelope (such as its rate of onset, or rise time, and to the depth of amplitude modulation) [11]. The envelopes of the bandpass outputs are formed by rectification and lowpass filtering [12]. The envelopes derived from each band was used to modulate white noise [13], as shown in Figure 1.

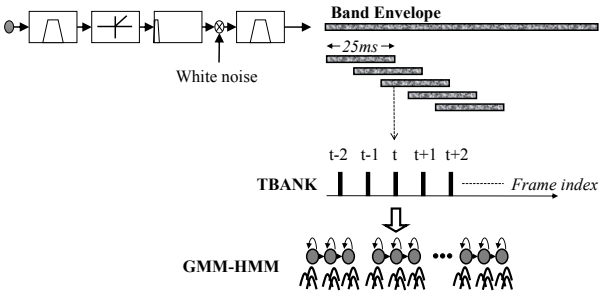


Figure 1: *Temporal alignment for supervised training of DNN.*

One problem in speech recognition is the problem of time [14]. To take into account the time distortions that may appear within its boundaries, a word is generally modeled by a sequence of states that can have variable time durations [1]. The HMM, based on dynamic programming operations, is a convenient tool to help port the strength of a static classifier to handle dynamic or sequential patterns [2]. The models are generally first trained using fixed alignments as targets (acoustic frames with the corresponding HMM state labels) [15]. Since the context dependent (CD)-DNN-HMM shares the phoneme tying structure and the HMM with the GMM-HMM system, the first step in the CD-DNN-HMM training is to train a GMM-HMM system using the training data [5]. After pre-training the DNN is fine-tuned using back-propagation with state labels obtained through forced alignment by maximum likelihood trained GMM-HMM models [16]. Since a GMM-HMM baseline creates the initial training labels for the DNN, it is important to have a good baseline system [7]. The decision tree used to cluster tri-phone states is also built using the GMM-HMM [5]. These alignments are often obtained from the forced-alignment of the supervised transcript with the acoustic frames using a GMM-HMM and can be further refined by realigning with a fully trained neural network and then by retraining the network with the new target alignments [15].

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

II. Flat Start Procedure

In order to build a set of HMMs, a set of speech data files and their associated transcriptions are required [17]. Initially such a state sequence has no timing information and we do not know which parts of the acoustic signal correspond to which states in the state sequence [3]. When no bootstrap data is available, a so-called *flat start* can be used [17]. This flat start procedure implies that during the first cycle of embedded re-estimation, each training utterance will be uniformly segmented [17]. It reads in a prototype HMM definition and some training data and outputs a new definition in which every mean and covariance is equal to the global speech mean and covariance [17], as shown in Figure 2.

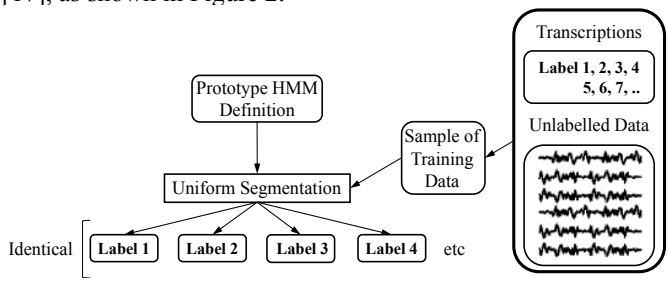


Figure 2: *Initialization with uniform segmentation of data*.

With the constraint of the state sequence for each utterance, such a model can generate a crude alignment which is sufficiently accurate for the EM algorithm to build upon, eventually producing an accurate speech recognition system [3]. The hope then is that enough of the phone models align with actual realizations of that phone so that on the second and subsequent iterations, the models align as intended [17].

III. Temporal Structure of HMM

In the traditional formulation of the HMM, individual states are assumed to be stationary stochastic sequences [18]. In each state there is no temporal information, and the path randomly moves around the mean, then exits on to the next state, arriving anywhere in the next state space [19]. This implies that within each state the speech vectors are associated with identical probability density functions which have the same mean and covariance [19], as shown in Figure 3.

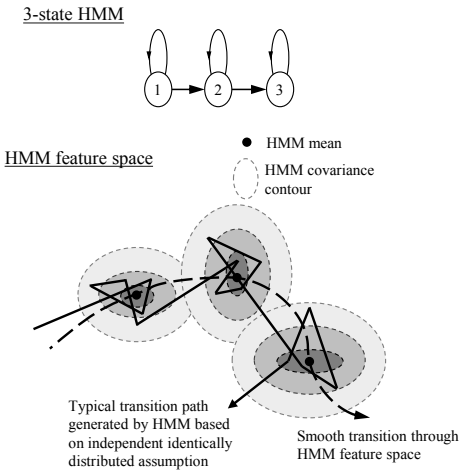


Figure 3: *Transition through the HMM feature space*.

IV. Frame Independence Assumption

Neighboring frames are highly correlated because of large overlaps in speech analysis windows [16]. Standard HMMs, based on the state-conditioned independent and identically distributed assumption, are known to be weak in capturing such correlations [20]. As time progresses the trajectory passes smoothly through the model [19], exiting each state near to the boundary of the next state. The strength of data correlations in the HMM source decays exponentially with time due to the Markov property, while the dependence among speech events does not follow such a fast and regular attenuation [20].

The actual features or abstractions learned by the network should be invariant under translation in time [9]. The message content must be retrieved from speech in a wide variety of listening conditions, including different talkers, environments, and amounts of distortions [13]. It is obvious that by including a long window of frames, the DNN model can exploit information in the neighboring frames [5]. Training the neural network to predict a distribution over senones causes more bits of information to be present in the neural network training labels [6]. Good models should provide correlation structures rich enough to accommodate the context dependence and other types of temporal dependence in speech data [20].

V. Experiments

A series of of experiments was performed on Aurora-4 [21], 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) [22]. The resulting features were further de-correlated using maximum likelihood linear transform (MLLT). 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 all trained and tested using 40-dimensional log mel filterbank (Fbank) 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.

VI. Simplified Representation

The MFCCs do not contain all spectral information in the speech signal [23]. One of the major problems with the cepstral features are that they are very sensitive to additive noise distortions [24]. The spectral information thus extracted may sometimes be degraded or reduced by certain inevitable factors like the limited bandwidth of a transmission channel or by the background noise [23]. Including some temporal information into the speech feature can lessen the effect of this assumption that speech is a stationary independent process, and can be used to improve recognition performance [19].

Work on prosthetic electrical stimulation of the auditory system by cochlear implants has re-focused attention on amplitude and temporal cues, which are the principal cues transmitted by these prostheses [13]. The presented results demonstrated that a simplified representation of speech is able to support relatively high levels of open-set recognition [12].

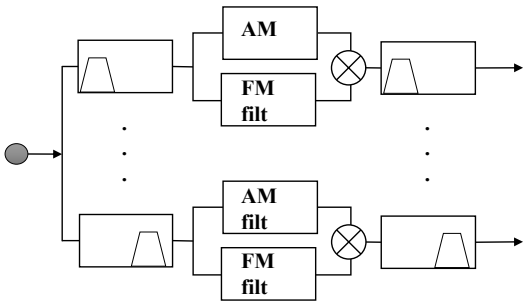


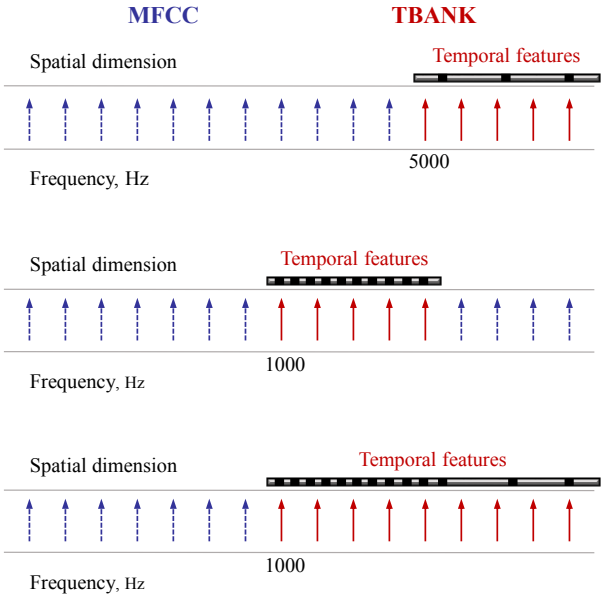
Figure 4: *Frequency amplitude modulation encoding (FAME).*

The FAME strategy [25], which proposes to extract the short varying FM components and integrates them in the cochlear implant, helps in reducing the F0 distortion in the short varying AM+FM spectrally reduced speech, compared to the AM-only spectrally reduced speech [26]. In the AM+FM condition, the FM is smoothed in terms of both rate and depth and then modulated by the AM [25]. The AM+FM stimuli were obtained by additionally frequency modulating each band's center frequency before amplitude modulation and subband summation, as shown in Figure 4. The modulation in speech, especially the FM, is expected to carry speaker-specific information [26]. The F0 normalized mean-square error (NMSE) analysis [26] performed on the cochlear implant-like spectrally reduced speech is a qualitative evidence supporting the speaker recognition subjective tests performed in [25].

The "slow" FM used here tracks gradual changes around a fixed frequency in the subband [25]. Here, we are not suggesting to replace the cepstral features (which have been a great success in the past) by the new features [24]. FM components support human speech recognition but make no significant improvement in ASR above a certain spectral resolution (8 subbands) [23]. Instead we want the new features to be used along with the cepstral features [24]. The coding efficiency can be improved, particularly for high-frequency bands, because the required sampling rate would be much lower to encode intensity and frequency changes with several hundred Hertz bandwidths than to encode, for example, a high-frequency subband at 5,000 Hz [25]. If this increases the dimensionality of the feature space, LDA may be used for dimensionality reduction [24].

VII. Results

Table 1 shows word error rate (WER%) for GMM-HMMs when trained and tested on alternative TBANK features. The context dependency trees is constructed using an alignment generated with a GMM-HMM [3]. These models were then used to align the training data to create senone labels for training the DNN. Table 1 shows DNNs give fewer errors if trained and tested on Fbank features after temporal alignment.



Tree-building Features	WER% (GMM)	WER% (DNN)
MFCC	5.08	2.88
+FAME (high)	4.76	2.45
+FAME (mid)	4.82	2.52
+FAME (mid+high)	4.67	2.54

Table 1: *Combining temporal feature representation at mid- and high-frequency regions during state-level alignment.*

Table 1 shows using a better alignment to generate training labels for the DNN improved the accuracy. Furthermore, the poorly matched alignment had a negative impact on the resulting system accuracy. Compared to previous [11] results (Fig. 1), higher accuracy was achieved in a balanced database [20] and by using FM (Fig. 4) while performing LDA+MLLT instead of using first and second-order derivatives.

The basic idea is to use the forced alignment to obtain frame-level senone labels for training the DNN and to borrow the triphone tying structure and transition probabilities from the CD-GMM-HMMs [27]. Three factors, in addition to the use of the DNN, contribute to the success: the use of tied triphones as the DNN modeling units, the use of the best available triphone GMM-HMM to generate the triphone state alignment, and the effective exploitation of a long window of input features [2]. 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 [6].

VIII. Effects of Temporal Alignment

Tree-building Features	(GMM) Del, Sub, Ins	(DNN) Del, Sub, Ins
MFCC	19, 189, 64	13, 114, 27
+FAME (high)	24, 180, 51	17, 96, 18
+FAME (mid)	22, 184, 52	15, 91, 29
+FAME (mid+high)	26, 182, 42	20, 90, 26

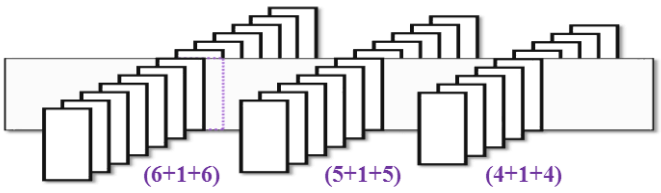
Table 2: *Error type (deletion, substitution, insertion) analysis*

Table 2 shows the lower the number of substitution errors in the system used during forced alignment to generate frame-level training labels for the neural net, the lower the error rate of the final neural-net-based system. Misclassification leads to substitution errors [28]. Previous optimization procedures explicitly attempted to minimize the number of word substitutions; this approach represents a move towards systems in which the training objective is maximum word accuracy [1].

If the segmentation process is not perfect, some sounds may be completely missed, which produces deletion errors [28]. The segmenter may also insert extra segment boundaries, which produces insertion errors [28]. By reducing extra segment boundaries, Table 2 shows how FM extraction can serve as front-end processing to help solve, automatically, the segregating and binding problem [25]. Moreover, the parameters adjustment of the analysis filter bank or the lowpass filters in the FAME strategy can affect the synthesized spectrally reduced speech's recognition results [23].

IX. Effects of Context Window

Table 6 shows the gain of DNN is almost entirely attributed to DNN's feature vectors that are concatenated from several consecutive speech frames within a relatively long context window [16]. In the standard DNN setup, each feature vector is augmented with its neighboring 10 frames within a context window (5+1+5) to form a 11-frame vector as DNN input feature [16]. Temporal alignment allowed the input layer to reduce from 440 to 360 visible units. On the other hand, if the context window is extended to augment too many neighboring frames as DNN input features, performance can degrade (e.g. using 13 frames in [16]). Naturally, a number of alternative signal representations could be used as input, but have not been tried in this study [9].



Tree-building Features	**Context window**		
	13 (6+1+6)	**11 (5+1+5)**	**9 (4+1+4)**
MFCC	2.76	2.88	2.84
+FAME (high)	2.69	2.45	2.58

Table 6: *DNN performance (WER %) using various context windows of past and future frames as input features*.

The most obvious conclusion from the results is that including temporal information into the speech feature improves recognition performance [19]. The network should have the ability to represent relationships between events in time [9]. Since the DNN training labels are generated from the GMM-HMM system and the label quality can affect the performance of the DNN system, it is important to train a good GMM-HMM system as the seed model [5].

From the computational economy viewpoint, fixed time alignment is by far superior to adaptive time alignment [29]. Some preprocessing is required to segment the part of the utterance on which the NN is going to focus its discriminant attention; therefore, this approach is difficult to extend to continuous speech [1]. Much higher variances are observable in the Gaussian distribution associated with the GMM states near the phoneme boundaries than with those near the center [30]. Performance of DNN is significantly improved when using a longer (5+1+5) context window [16]. Contrary to common sense, however, by including neighboring frames the DNN also models correlations between frames of features and thus partially alleviates the violation to the observation independence assumption made in HMMs [5].

X. Discussion

It is expected that the performance gap between acoustic models that use DNNs and ones that use GMMs will continue to increase for some time [7]. However, increasing the fine-tuning (labeled) data is much more important than increasing the pre-training (unlabeled) data [27]. Therefore, the presented temporal alignment results reveal the self-contradictory nature of performance gap expectations [7] when one of three factors that contribute to the success is the use of the best available triphone GMM-HMM to generate the senone alignment [27].

While DNNs have become the dominant acoustic model for speech recognition systems, they are still dependent on GMMs for alignments both for supervised training and for context dependent tree building [3]. By building the CD trees on an alignment from a DNN and using features better suited to a DNN, it was shown that such GMM-free training [3] can result in better models than when using conventional, GMM-based flat starting and CD tree building. The advantage of this approach is that the state inventory is matched to the DNN model as opposed to the state inventory from the GMM which is mismatched in terms of the features and model family used to design it [8].

XI. Conclusions

Time alignment presents the greatest problem for DNN based systems defined in terms of static pattern classification tasks. Alternative features were introduced for the investigation of the impact of iteratively realigning with larger, state-of-the art models from flat start context independent alignments. Future work will address combining TBANK features as DNN input.

XII. Acknowledgments

This research was supported by National Science Council of Taiwan under contracts NSC104-2221-E-001-026-MY2.

REFERENCES

[1] P. Haffner, M. Franzini, and A. Waibel, "Integrating time alignment and neural networks for high performance continuous speech recognition," *in Proc. ICASSP*, 1991, pp. 105-108.

[2] L. Deng, and D. Yu, "Deep learning: methods and applications," *in Foundations and trends in signal processing,* 7 (2-4), pp. 197-387, 2013.

[3] A. Senior, G. Heigold, M. Bacchiani, and H. Liao, "GMM-free DNN training," *in Proc. ICASSP*, 2014.

[4] M. Franzini, K.-F. Lee, and A. Waibel, "Connectionist Viterbi training: a new hybrid method for continuous speech recognition," *in Proc. ICASSP*, 1990, pp. 425-428.

[5] D. Yu, and L. Deng, "Deep neural network-hidden Markov model hybrid systems," *in Automatic Speech Recognition*, Springer London, 2015, pp. 99-116.

[6] 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.

[7] 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.

[8] M. Bacchiani, A. Senior, and G. Heigold, "Asynchronous, online, GMM-free training of a context dependent acoustic model for speech recognition," *in Proc. Interspeech,* 2014.

[9] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoust. Speech Signal Process.*, 37 (3), pp. 328-339, 1989.

[10] F. Jelinek, L.R. Bahl, and R.L. Mercer, "Design of a linguistic statistical decoder for the recognition of continuous speech," *IEEE Trans. Information Theory*, pp. 250-156, 1975.

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

[12] 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.

[13] 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.

[14] A. Waibel, "Modular construction of time-delay neural networks for speech recognition," *Neural Computation*, 1 (1), pp. 39-46, 1989.

[15] H. Sak, A. Senior, K. Rao, O. Irsoy, A. Graves, F. Beaufays, and J. Schalkwyk, "Learning acoustic frame labeling for speech recognition with recurrent neural networks," *in Proc. ICASSP*, 2015.

[16] J. Pan, C. Liu, Z. Wang, Y. Hu, and H. Jiang, "Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: why DNN surpasses GMMs in acoustic modeling*," in Proc. ISCSLP,* 2012, pp. 301-305.

[17] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book," 3.4 edition, 2009

[18] L. Deng, M. Aksmanovic, X. Sun, and J. Wu, "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states," *IEEE Trans. Speech, Audio, Process.*, 2 (4), pp. 507-520, 1994.

[19] B. Milner, "Inclusion of temporal information into features for speech recognition," in *Proc. ICSLP*, 1, pp. 256-259, 1996.

[20] L. Deng, K. Hassanein, and M. Elmasry, "Neural-network architecture for linear and nonlinear predictive hidden Markov models: applications to speech recognition," *Proc. IEEE Workshop on Neural Networks for Sig. Process.*, pp. 411-422, 1991.

[21] N. Parihar, J. Picone, D. Pearce, and H.G. Hirsch, "Performance analysis of the Aurora large vocabulary baseline system," *in Proc. EUSIPCO*, 2004, pp. 553-556.

[22] R.O. Duda, P.E. Hart, and D.G. Stork, "Pattern classification," *in Wiley*, November 2000.

[23] C.-T. Do, M.V. Kumar, D. Pastor, and A. Goalic, "Automatic speech recognition of cochlear implant-like spectrally reduced speech," *in NCC,* 2009.

[24] K. K. Paliwal, "Spectral subband centroid features for speech recognition," *in Proc. ICASSP*, 1998, pp. 617-620.

[25] F.-G. Zeng, K. Nie, G.S. Stickney, Y.Y. Kong, M. Vongphoe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *in Proc. Nat. Acad. Sci. USA (PNAS)*, 102 (7), pp. 2293-2298, 2005.

[26] C.-T. Do, D. Pastor, and A. Goalic, "On normalized MSE analysis of speech fundamental frequency in the cochlear implant-like spectrally reduced speech," *IEEE Trans. Biomed. Eng.*, 57 (3), 2010.

[27] D. Yu, L. Deng, and G.E. Dahl, "Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition," *in Proc. NIPS Workshop on Deep Learning*, 2010.

[28] L.R. Bahl, and F. Jelinek, "Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition," *IEEE Trans Information Theory*, 21 (4), pp. 404-411, 1975.

[29] H. Sakoe, R. Isotani, K. Yoshida, K. Iso, and T. Watanabe, "Speaker-independent word recognition using dynamic programming neural networks," *in Proc. ICASSP*, 1989, pp. 29-32.

[30] L. Deng, P. Kenny, M. Lennig, V. Gupta, F. Seitz, and P. Mermelstein, "Phonemic hidden Markov models with continuous mixture output densities for large vocabulary word recognition," *IEEE Trans. Speech, Audio, Process.*, 39 (7), pp. 1677-1681, 1991.