ROBUST EXCITATION-BASED FEATURES FOR AUTOMATIC SPEECH RECOGNITION

Thomas Drugman[1], *Yannis Stylianou*[1], *Langzhou Chen*[1], *Xie Chen*[2], *Mark J.F. Gales*[2]

1. Toshiba Research Europe Ltd., Cambridge Research Lab, Cambridge, U.K.
2. University of Cambridge Engineering Dept, Trumpington St., Cambridge, U.K.

ABSTRACT

In this paper we investigate the use of noise-robust features characterizing the speech excitation signal as complementary features to the usually considered vocal tract based features for Automatic Speech Recognition (ASR). The proposed Excitation-based Features (EBF) are tested in a state-of-the-art Deep Neural Network (DNN) based hybrid acoustic model for speech recognition. The suggested excitation features expand the set of periodicity features previously considered for ASR, expecting that these features help in a better discrimination of the broad phonetic classes (e.g., fricatives, nasal, vowels, etc.). Our experiments on the AMI meeting transcription system showed that the proposed EBF yield a relative word error rate reduction of about 5% when combined with conventional PLP features. Further experiments led on Aurora4 confirmed the robustness of the EBF to both additive and convolutive noises, with a relative improvement of 4.3% obtained by combinining them with mel filter banks.

Index Terms— neural networks, automatic speech recognition, speech excitation signal

1. INTRODUCTION

Recent and promising studies in Deep Neural Networks (DNN) have shown [1] that they have the ability to clearly outperform the standard Gaussian Mixture Model (GMM) approach for acoustic modeling in Automatic Speech Recognition (ASR). Most of the works in feature extraction over the past decades were however carried out in the frame of GMM-based modeling, and features were designed specifically in that context.

The two most popular feature extraction schemes are probably the Mel Frequency Cepstral Coefficients (MFCCs, [2]) and the Perceptual Linear Prediction (PLP, [3]) features. Recently, the Power Normalized Cepstral Coefficients (PNCCs, [4]) have also received a particular attention due to the robustness of their performance in GMM-based acoustic modeling. Various other types of features have been proposed in the literature. Some are based on auditory models (e.g. [5]). Some others aim at replacing the power Fourier spectrum by alternative representations of the vocal tract response. These include the Minimum Variance Distortionless Response (MVDR, [6]) or Group Delay-based features [7, 8].

All these features have been relatively extensively studied in ASR systems based on the use of Hidden Markov Models (HMM) coupled with GMM. However recent progress has been made in the use of DNNs to model the HMM state posteriors. These advances have also opened new perspectives in feature extraction. The use of DNNs indeed does not imply any assumption about the correlation between the features or about the Gaussianity of their distributions. Our preliminary experiments indicated that features which were designed for robust GMM-based ASR no longer outperform simple features such as Mel-log filter banks (FBANK). Moreover, they also showed that combinations between these features do not bring any significant improvement in ASR (if any). We believe that this is because the great majority of feature extraction schemes rely on a representation of the same information: the vocal tract filter. The focus has therefore now moved towards finding features which are complementary with spectral envelope-based representations.

Very few studies have focused on the use of excitation-based features for ASR. The first attempt was made by Thomson [9, 10] who proposed the use of two voicing measures: an auto-correlation based measure of periodicity and the jitter to characterize the inter-frame pitch variation. When combined to cepstral features, a relative reduction of 40% of the string error rate was obtained on a connected digit recogntion task. In [11], Zolnay et al. studied three different voicing features as additional acoustic features for continuous speech recognition. These features were extracted from the harmonic product spectrum, the autocorrelation and the average magnitude difference function. Relative improvements up to 6% were achieved on a large-vocabulary task relatively compared to using MFCCs alone. Finally, in [12], Ishizuka et al. proposed a method which decomposes the speech signal into periodic and nonperiodic components using comb filters independently designed in various subbands.

In this paper, we propose robust excitation-based features and investigate how they can be helpful in improving ASR performance on various databases. The set of already suggested features is expanded by considering robust pitch tracking algorithms, and voice quality measurements. Experiments are conducted on two databases, well established for noise-robust ASR: AMI meeting transcription system and Aurora

4. Results support the arguments that excitation-based features provide complementary information to the vocal tract-based features, while it is possible to extract these features in a robust way, even in very noisy environments as in the two databases we considered.

The paper is structured as follows. Section 2 describes the proposed robust excitation-based features. The results of our experiments are discussed in Section 3. Section 4 finally concludes the paper.

2. ROBUST EXCITATION-BASED FEATURES

According to the mechanism of voice production, speech is considered as the result of a glottal flow (also called *source* or *excitation* signal) filtered by the vocal tract cavities [13]. This led to the well-known *source-filter* model which motivates the present study: source and filter features reflect different physiological characteristics of speech. They are expected to be complementary, which could be turned into advantage in an ASR system.

Speech excitation usually refers to the glottal flow signal. The glottal flow has been already shown to be useful in various speech processing applications [14, 13]. However these works were conducted in relatively well-controlled situations in which the detrimental effects of the noise are quite limited. A reliable and accurate estimation of the glottal flow in adverse conditions is still an open and challenging problem [15]. Nevertheless, it is possible to extract relevant features of the excitation signal without requiring an explicit estimation of the glottal flow. We now describe the proposed Excitation-based features (EBF) which we will use for our ASR experiments.

Excitation-based features can be extracted in the time, the frequency or the cepstral domain. They can also be computed directly from the speech signal, or from the Linear Prediciton (LP) residual signal, obtained by inverse filtering after removing the contribution of the spectral envelope. The advantage of working with the LP residual is that it exhibits relevant characteristics of the glottal source [13] while circumventing complex and noise-sensitive operations (e.g. pitch-synchronous analysis) involved in the majority of glottal flow estimation techniques [13, 15].

In the time domain, a popular and very simple periodicity feature is the zero-crossing rate (ZCR) which indirectly measures the degree of voicing from the speech signal. Another common approach to quantify periodicity relies on the auto-correlation (AC) function of the speech signal [9, 11] by measuring the relative height of the maximum of this function in the plausible pitch range. The Average Magnitude Difference Function (AMDF) can be formulated as a function of the AC function. The relative depth of the minimum AMDF valley in the plausible pitch range has been used for ASR in [11] and Voice Activity Detection (VAD) in [16]. The normalized LP error was proposed in [17] for VAD purpose. It quantifies how well an auto-regressive model fits the signal, and lower errors are expected in voiced sounds. Finally, high-order statistics of the LP residual have also been proposed in the literature [17, 18]. The kurtosis of the LP residual has been used for VAD purpose in [17] and as a measure of the sparsity of the excitation in [19] to characterize the discontinuities at the glottal closure instants. As for the skewness of the residue, it captures the asymmetry of the excitation and is related to the polarity of the speech signal [18].

In the spectral domain, the Harmonic Product Spectrum (HPS), defined as the product of R frequency-shrunken replicas of the speech amplitude spectrum, has been proposed for ASR and VAD respctively in [11] and [16]. A HPS-based periodicity measure consists of the maximum HPS peak in the plausible pitch range. We also employ two features extracted from the Summation of the Residual Harmonics (SRH) algorithm [20], which was shown to be one of the most robust pitch tracker. This method is based on the spectrum $E(f)$ of the residual excitation and the SRH value is computed as:

$$SRH = \underset{f}{\operatorname{argmax}}(E(f) + \sum_{k=2}^{N_{harm}} [E(k \cdot f) - E((k - \frac{1}{2}) \cdot f)]),$$

where the number of harmonics N_{harm} is fixed to 5 as in [20], and where f is varied in the plausible pitch range. SRH criterion differs from HPS in mainly two aspects: *i)* it exploits the residual signal, which allows to minimize the effects of both the vocal tract resonance and of the noise [20], *ii)* it involves also interharmonics. The two features used in this work differ by the energy-normalization or not of $E(f)$ for each frame.

Finally, as cepstral-domain feature, the Cepstral Peak Prominence (CPP) was originally proposed in [21] for the prediction of breathiness ratings. CPP is a measure of the amplitude of the cepstral peak corresponding to the fundamental period, normalized for overall signal amplitude.

In total, 10 excitation-based features are considered in the rest of this paper, and they will be referred to as EBF: the ZCR, the height of the AC function, the depth of the AMDF, the normalized LP error, the kurtosis and the skewness of the LP residual, the maximum of the HPS, the 2 SRH-based measurements and CPP. In all cases, the plausible pitch range is fixed to $[60 - 400]Hz$. All implementations are conform to the descriptions provided in the original publications. Note that the implementations of CPP and SRH are available from the COVAREP project [22].

3. EXPERIMENTAL RESULTS

In a first experiment, we investigated which of the proposed EBF were the most relevant by assessing their class separability (Section 3.1). The usefulness of the proposed EBF was then studied for two very different ASR tasks. One was based

on the Augmented Multi-party Interaction (AMI) data set, i.e. AMI meeting transcription task (Section 3.2). This is a multi-accent, spontaneous speech recognition task with large training data and large vocabulary. The other is the Aurora 4 noise robust speech recognition task with small training data and medium vocabulary size (Section 3.3).

3.1. Class Separability of the proposed features

Before any use in a complete ASR system, we first performed an objective assessment of the proposed EBF in terms of their intrinsic discrimination power. For this purpose, the metric we used is Fisher's class separability. This measure is used in Linear Discriminant Analysis (LDA) and considers the ratio of the variance between the classes to the variance within the classes. Phones are here used as classes. The phone boundaries were obtained by forced alignment using the Train&Align tool [23]. As speech material, we used 3000 randomly-chosen utterances from the Aurora4 database (see Section 3.3) in clean conditions. We also simulated noisy conditions by artificially adding a babble noise (taken from the Noisex92 database) at 0 dB SNR.

The results are reported in Figure 1 and reflect the relative improvement in class separability brought by each feature individually when combined with 13-dimensional conventional PLP features. Except for the residual skewness, the proposed EBF add a sensible increase of class separability varying from 1.3% to 7.5% in clean conditions. A similar observation holds in the noisy condition, where ZCR even reaches a relative improvement of 8.6% alone. When the 10 proposed EBF are combined with PLP, the total increase of class separability is respectively of 22.8 and 27.3% in clean and noisy conditions. These first results show evidence that the proposed EBF have the potential to enhance ASR performance both in controlled and adversed environments.

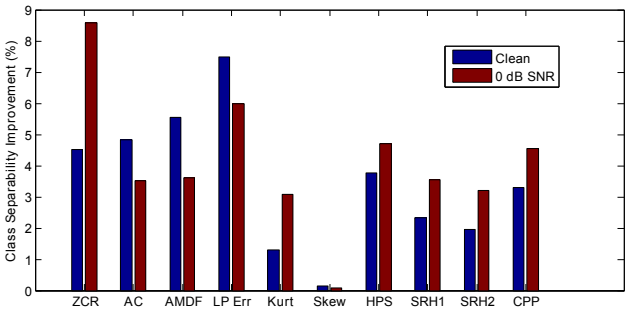

Fig. 1. Relative improvement (in %) in class separability brought by each of the proposed EBF.

3.2. AMI Meeting Transcription Experiments

Our first ASR experiment was carried out on the AMI corpus. This corpus [24] was collected for research and development of technology that will help groups interact better. As part of this corpus close-talking and far-field microphones with high quality transcriptions are available. Numerous previous studies have reported results on this corpus [25, 26, 27, 28]. In this work only the far-field microphones, multiple distant microphone data (MDM) was used. Additionally overlapping speech data was removed. This yielded about 59 hours of data. In addition to the AMI corpus, 52 hours from the ICSI corpus [29] and 10 hours from the NIST corpus [30] were used. ICSI meeting data was recorded in the conference room in ICSI. Beamforming is performed using the *BeamformIt* tool [31] to yield a single audio channel [1].

Four meetings are held back from the AMI data to give an AMI dev and eval sets, each with two sets of meetings and 4 speakers per meeting. As overlapping speech is not evaluated this yielded a total test set duration of about 5.29 hours. The total available data for training, after removing the 4 meetings is about 121 hour-long data. This is the same configuration and held-out test sets as those used in [26]. Automatic segmentation is used for evaluation.

The acoustic models based on hybrid systems were constructed as follows. A DNN with four hidden layers and 1000 nodes per layer was trained. Nine consecutive frames were concatenated as input features of the DNN. This latter was trained in a supervised and discriminative fashion layer by layer in pretraining [33], followed by a fine-tuning with several epochs until the frame accuracy converges in the cross-validation set. The alignment for the targets was obtained from a well-trained Speaker Adaptive Training (SAT) Tandem system. 6000 distinct states were clustered from the decision tree in the GMM-HMM system, which were further used as targets in the training of DNN. Two sets of basic features were used: 13-dimensional PLP and 26-dimensional mel filter banks (FBANK), together with their first, second and triple deltas appended. Another two sets of compound features were constructed by concatenating the 10-dimensional EBF with PLP or FBANK features, again with first, second and triple deltas appended. Cepstral mean and variance normalization at the speaker level were applied to all features before being fed into the DNN.

The 3-gram language model used in this paper is the same as the one used in [26]. These used a 41K word-list and were trained on a variety of sources including the AMI, ICSI, NIST and ISL corpora transcriptions, Callhome, Switchboard, Gigaword and web data collected by the University of Washington. Language model interpolation weights were tuned on the AMI dev set. In total, 2.5G words of language model training data were used.

Table 1 gives the experimental results of the speaker-independent (SI) hybrid system. The Word Error Rate (WER) from the output of confusion network decoding is reported. Two main conclusions can be drawn from these results: *i)* FBANK consistently outperforms PLP; *ii)* an improvement is

[1] Currently there is no Wiener filtering in the front-end processing, as used for example in [32], which should yield performance gains.

obtained by combining the proposed EBF to conventional vocal tract-based features. An average relative WER reduction of respectively 4.77% and 1.64% is achieved when EBF are used jointly with PLP and FBANK.

Table 1. WER results (in %) of on the AMI corpus

MLP feature	WER	
	dev	eval
PLP	35.7	35.6
+EBF	34.1	33.8
FBANK	34.0	33.0
+EBF	33.3	32.6

3.3. Aurora 4 Experiments

In our second ASR experiment, the use of the proposed EBF was investigated in the Aurora 4 task. This is a noise-robust continuous speech recognition task with a size of vocabulary of 5k. The Aurora 4 database is from WSJ data set in which the additive noise and convolutional distortion has been artificially added. Two training sets were defined: the clean training set and the multi-condition training set. The clean set includes 7138 utterances recorded by the primary Sennheiser microphone. The multi-conditional training set consists of the same utterances from the primary Sennheiser microphone and from a secondary microphone which includes convolutional distortions. The multi-condition training set includes clean condition and 6 noise conditions, i.e. airport, babble, car, restaurant, street and train station. The Aurora 4 test data consists of 330 utterances from 8 speakers, recorded by the same two channels under the same clean and 6 noisy conditions as in the training data. This thus leads to a total of 14 test sets.

In this work the multi-condition training set was used for system training. As conventional vocal tract-based features, we used 25-dimensional FBANK as they provided the best performance in Section 3.2. The static feature vectors were spliced in time taking a context of ± 3 frames. Then the linear discriminant analysis (LDA) was used to reduce the dimension of the spliced features from 175 to 75. It was followed by a global semi-tied covariance (STC) matrix for de-correlation. The DNN hybrid system with SAT was used as an acoustic model. For each speaker and noise condition, a global Constrained Maximum Likelihood Linear Regression (CMLLR)transform was trained and cascaded with the LDA+STC transforms to account for speaker and noise variability. This transformed feature vector was concatenated with the proposed EBF as input to the DNN. Again, the features were spliced in time with a window of ± 5 frames. It was followed by a global mean and variance normalization. The DNN used in this work contains 4 hidden layers and 2000 nodes for each hidden layer. The alignments for the target output were from a SAT based GMM-HMM system with about 3k tied context dependent states. The Deep Belief Network (DBN) based pre-training was used to initialize the DNN. Both cross-entropy (XEnt based training and the Segmental Minimum Bayes Risk (SMBR) based sequence training were used for fine-tuning. The results are given in Table 2.

Table 2. WER (in %) results for Aurora 4

channel	noise	FBANK		FBANK+EBF	
		XEnt	SMBR	XEnt	SMBR
1	clean	3.72	3.70	3.87	3.75
	airport	5.81	5.49	6.07	5.55
	babble	6.02	5.47	6.13	5.51
	car	4.28	4.24	4.24	4.15
	restaurant	8.43	7.83	8.03	7.62
	street	8.03	7.08	7.92	6.80
	train	7.29	6.78	7.38	6.63
2	clean	6.58	6.05	5.57	4.89
	airport	17.62	16.33	16.93	14.89
	babble	18.14	17.04	17.65	16.20
	car	9.47	8.52	8.09	7.36
	restaurant	20.98	19.71	21.54	20.23
	street	20.33	18.91	20.53	18.53
	train	20.13	18.65	19.72	17.44
avg.		11.20	10.41	10.98	9.96

Three main observations can be drawn from Table 2: *i)* SMBR clearly outperforms XEnt as training method for fine-tuning; *ii)* Corroborating the results from Section 3.2 on a very different task, the proposed EBF yield also an improvement when combined with FBANK. The best WER obtained is 9.96% and is possible thanks to a relative reduction of 4.3% when using EBF in complement with FBANK; *iii)* Interestingly, the proposed EBF are seen to be on overall helpful in both additive and convolutive (channel 1 vs. channel 2) noise, which further supports their noise robustness. Finally, it is worth noting that our attempts to combine MFCC, PLP or PNCC with FBANK did not result in any relevant improvement (when any) which confirms the need to look for complemenatry features, such as the proposed EBF.

4. CONCLUSION

This work investigated the use of excitation-based features to complement conventional vocal tract-based acoustic features to improve the performance DNN-based ASR systems. Ten robust excitation-based features (EBF) were proposed and were evaluated on two very different ASR tasks: the AMI meeting transcription and Aurora 4. The experimental results showed that the proposed EBF provides a relative improvement varying between 1.6 and 4.8% when they were combined with standard PLP or FBANK features. Furthermore, this improvement was observed consistently for the two ASR tasks and across both additive and convolutive noisy conditions.

5. REFERENCES

[1] G. Hinton, L. Deng, D. Yu, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, G. Dahl, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," in *IEEE Signal Processing Magazine*, 2012, vol. 29(6), pp. 82–97.

[2] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *IEEE. Trans. on Acoustics, Speech, and Signal Processing*, 1980, vol. 28(4), pp. 357–366.

[3] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," in *Journal of the Acoustical Society of America*, 1990, vol. 87(4), pp. 1738–1752.

[4] C. Kim and R. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *IEEE Conf. on Acoustics, Speech and Signal Processing*, 2012, pp. 4101–4104.

[5] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *IEEE Conf. on Acoustics, Speech and Signal Processing*, 2009, pp. 4625–4628.

[6] M. Alam, P. Kenny, and D. O'Shaughnessy, "Speech recognition using regularized minimum variance distortionless response spectrum estimation-based cepstral features," in *IEEE Conf. on Acoustics, Speech and Signal Processing*, 2013, pp. 8071–8075.

[7] B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp group delay analysis of speech signals," in *Speech Communication*, 2007, vol. 49(3), pp. 159–176.

[8] E. Loweimi, S. Ahadi, and T. Drugman, "A new phase-based feature representation for robust speech recognition," in *IEEE Conf. on Acoustics, Speech and Signal Processing*, 2013, pp. 7155–7159.

[9] D. Thomson and R. Chengalvarayan, "Use of periodicity and jitter as speech recognition features," in *IEEE Conf. on Acoustics, Speech and Signal Processing*, 1998, pp. 21–24.

[10] D. Thomson and R. Chengalvarayan, "Use of voicing features in hmm-based speech recognition," in *Speech Communication*, 2002, vol. 37, pp. 197–211.

[11] A. Zolnay, R. Schluter, and H. Ney, "Extraction methods of voicing feature for robust speech recognition," in *Proc. Eurospeech*, 2003, pp. 497–500.

[12] K. Ishizuka, T. Nakatani, Y. Minami, and N. Miyazaki, "Speech feature extraction method using subband-based periodicity and nonperiodicity decomposition," in *Journal of the Acoustical Society of America*, 2006, vol. 120(1), pp. 443–452.

[13] T. Drugman, P. Alku, B. Yegnanarayana, and A. Alwan, "Glottal source processing: from analysis to applications," in *Computer Speech and Language*, 2014, vol. 28(5), pp. 1117–1138.

[14] T. Drugman, "Advances in glottal analysis and its applications," in *PhD thesis, University of Mons*, 2011.

[15] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," in *Computer Speech and Language*, 2012, vol. 26(1), pp. 20–34.

[16] S. Sadjadi and J. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," in *IEEE Sig. Pro. Letters*, 2013, vol. 20, pp. 197–20.

[17] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the lpc residual domain," in *IEEE Trans. Speech Audio Process.*, 2001, vol. 9, pp. 217–231.

[18] T. Drugman, "Residual excitation skewness for automatic speech polarity detection," in *IEEE Signal Processing Letters*, 2013, vol. 20(4), pp. 387–390.

[19] T. Drugman, "Maximum phase modeling for sparse linear prediction of speech," in *IEEE Signal Processing Letters*, 2014, vol. 21(2), pp. 185–189.

[20] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. Interspeech*, 2011, pp. 1973–1976.

[21] J. Hillenbrand and R. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," in *Journal of Speech and Hearing Research*, 1996, vol. 39, pp. 311–321.

[22] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep - a collaborative voice analysis repository for speech technologies," in *IEEE Conf. on Acoustics, Speech and Signal Processing*, 2014, pp. 960–964.

[23] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "Train&align: A new online tool for automatic phonetic alignment," in *IEEE Spoken Language Technology Workshop*, 2012, pp. 416–421.

[24] J. Carletta et al., "The AMI meeting corpus: A pre-announcement," in *Machine learning for multimodal interaction*, pp. 28–39. Springer, 2006.

[25] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings the AMI and AMIDA projects," in *ASRU, IEEE Workshop on*. IEEE, 2007, pp. 238–247.

[26] C. Breslin, KK Chin, M. Gales, and K. Knill, "Integrated online speaker clustering and adaptation.," in *Proc. ISCA Interspeech*, 2011, pp. 1085–1088.

[27] T. Hain, L. Burget, J. Dines, P. Garner, F. Grezl, A. Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "Transcribing meetings with the AMIDA systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 486–498, 2012.

[28] X. Chen, M. Gales, K. Knill, C. Breslin, L. Chen, K.K. Chin, and Vincent Wan, "An initial investigation of long-term adaptation for meeting transcription," in *Proc. INTERSPEECH*, 2014.

[29] Ad. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, et al., "The ICSI meeting corpus," in *Proc. ICASSP*. IEEE, 2003, vol. 1, pp. I–364.

[30] J. Garofolo, C. Laprun, M. Michel, V. Stanford, and E. Tabassi, "The NIST meeting room pilot corpus.," in *LREC*, 2004.

[31] Xavier Anguera Miro, *Robust speaker diarization for meetings*, Ph.D. thesis, 2007.

[32] T. Hain, V. Wan, L. Burget, M. Karafiat, J. Dines, J. Vepa, G. Garau, and M. Lincoln, "The AMI system for the transcription of speech in meetings," in *Proc. ICASSP*. IEEE, 2007, vol. 4, pp. IV–357.

[33] Frank Seide, Gang Li, Xie Chen, and Dong Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *ASRU, IEEE Workshop on*. IEEE, 2011, pp. 24–29.