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 robust to noise features characterizing the speech excitation signal as complementary features to the usually considered vocal tract based features for automatic speech recognition (ASR). The features are tested in a state-of-the-art Deep Neural Network (DNN) based hybrid acoustic model for speech recognition. The suggested excitation features expands the set of excitation 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.). Relative improvements in the word error rate are observed in the AMI meeting transcription system with greater gains (about 5%) if PLP features are combined with the suggested excitation features. For Aurora 4, significant improvements are observed as well. Combining the suggested excitation features with filter banks, a word error rate of 9.96% is achieved.

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

1. INTRODUCTION

The recent and promising advances in Deep Neural Network (DNN) based acoustic modeling have 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. Features which were recently designed for robust Gaussian Mixture Model (GMM) based speech recognition no longer outperform simple features such as Mel-log filter banks. Moreover, combinations between these features do not bring any significant improvement in ASR. 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.

For GMM-HMM based ASR, the two most popular feature extraction schemes are probably the Mel Frequency Cepstral Coefficients (MFCCs, [1]) and the Perceptual Linear Prediction (PLP, [2]) features. Recently, the Power Normalized Cepstral Coefficients (PNCCs, [3]) 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 perceptual considerations. 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, [4]) or Group Delay-based features [5, 6].

All aforementioned features characterize the same acoustic information: the spectral envelope which is mainly due to the vocal tract filter. The venue of DNN-based acoustic modeling opens new perspectives in the field of feature extraction, as the constraints on the distribution and correlation of the features are released. The focus has therefore now moved towards finding features which are complementary with spectral envelope-based representations.

Various few studies have focused on the use of excitation-based features for ASR. The first attempt was made by Thomson [7, 8] 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 [9], Zolnay et al. studied three different voicing features as additional acoustic features for continuous speech recognition. These features are 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 [10], 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 quality measurements of speech. Experiments are conducted on two databases, well established for noise robust ASR: AMI meeting transcription system and Aurora 4. Results supports 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 a 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 experimental protocol and the results of our experiments are discussed in Section 3. Section 4 finally concludes the paper.

2. DNN HYBRID SYSTEM WITH 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 [11]. 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.

In a DNN hybrid ASR system, the DNN is used to predict the posterior distribution of the context dependent HMM states defined by a traditional context dependent HMM model. The input of the DNN is the spliced acoustic feature vectors within a context window. The vocal tract based acoustic features, e.g. MFCC, PLP and filter bank (FBANK) are widely used as the input features of DNN. To normalize the speaker or environment factors, the linear transforms, e.g. constrained maximum likelihood linear regression (CMLLR) can be applied to the feature vectors. This yields the speaker adaptive training (SAT) system. This work introduces various excitation features into the DNN-HMM based ASR systems. These excitation features are concatenated with the traditional vocal track based features as the input of the DNN. The framework is shown in figure 1. In that figure, the CMLLR transform is optional, without it, a speaker independent system is constructed.

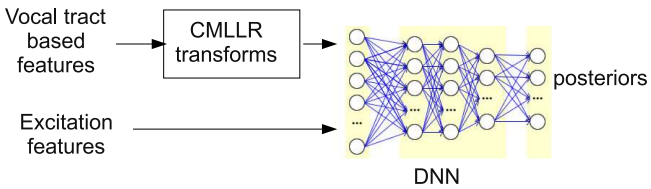

Fig. 1. *Hybrid system with excitation features*

2.1. Robust Excitation-based Features

In this work, various excitation features were investigated, and they are described in this section.

Speech excitation usually refers to the glottal flow signal. The glottal flow has been already shown to be useful in various speech processing applications [12, 11]. 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 [13]. Nevertheless, it is possible to extract relevant features of the excitation signal without requiring an explicit estimation of the glottal flow. This paper focuses on such parameters which can be used for robust ASR.

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 [11] while circumventing complex and noise-sensitive operations (e.g. pitch-synchronous analysis) involved in the majority of glottal flow estimation techniques [11, 13].

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 [7, 9] 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 [9] and VAD in [14]. The normalized LP error was proposed in [15] for VAD. 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 [15, 16]. The kurtosis of the LP residual has been used for Voice Activity Detection (VAD) purpose in [15] and as a measure of the sparsity of the excitation in [17] to characterize the discontinuities at the glottal closure instants.

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 [9] and [14]. 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 [18], 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 [18], 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 [18], *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 [19] 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, 9 excitation-based features are considered in the rest of this paper, and they will be referred to as EBF features: the ZCR, the height of the AC function, the depth of the AMDF, the normalized LP error, the residual kurtosis, 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 [20].

3. EXPERIMENTAL RESULTS

The excitation features proposed in this work was investigated in two very different ASR tasks. One was based on the Augmented Multi-party Interaction (AMI) data set, i.e. AMI meeting transcription task. This is a multi-accent, spontaneous speech recognition task with large training data and large vocabulary. The other is Aurora 4 noise robust speech recognition task with small training data and medium vocabulary size.

3.1. AMI meeting transcription experiments

The first part of experiments were based on AMI meeting transcription system. AMI corpus [21] 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. It was investigated in a number of previous work [22, 23, 24, 25]. 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 [26] and 10 hours from the NIST corpus were used [27]. ICSI meeting data was recorded in the conference room in ICSI. Beamformed is performed using the *BeamformIt* tool [28] to yield a single audio channel. [1].

Four meetings are held back from the AMI data to give an AMI dev and eval set, each with two sets of meetings and 4 speakers per meeting. As overlapping speech is not evaluated this yielded a total test set size of about 5.29 hours. The total available data for training, after removing the 4 meetings is about 121 hours of data. This is the same configuration, and held-out test sets, as used in [23]. Automatic segmentation is used for evaluation.

[1]Currently there is no Wiener filtering in the front-end processing, as used for example in [29], which should yield performance gains.

The acoustic models based on hybrid systems are constructed. A DNN with four hidden layers, with 1000 nodes per layer were trained. 9 consecutive frames were concatenated as input feature for deep neural network. The DNN was trained in a supervised fashion and discriminatively layer by layer in pretraining [30], and followed by fine-tune with several epochs until the frame accuracy converges in cross validation set. The alignment for the targets was obtained from a well-trained SAT Tandem system. 6000 distinct states were clustered from decision tree in GMM-HMM systems, which was used as target in the training of DNN, two sets of basic feature was used, 13-dimensional PLP and 26-dimensional FBANK, with their first, second, triple delta appended. Another two sets of compound feature were constructed by concatenating the 10-dimensional EBF feature with PLP or FBANK feature, again, with first, second and triple delta appended. Performance on these feature will be compared in Section 3. CMN and CVN on speaker and show level were performed on features before being fed into neural networks.

The 3-gram language model used in this paper is the same as used in [23]. 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 with EBF feature. The WER from the output of confusion network decoding is reported. According to the results, FBK consistently outperform PLP in hybrid systems. EBF feature helps to reduce WER on both PLP and FBK feature. The concatenated EBF feature gives 1.2% and 1.8% absolute (4.5% and 5.1% relatively) WER reduction. The effect of EBF on FBANK is smaller, giving 0.6% and 0.4% absolute improvement (2.0% and 1.2% relatively).

Table 1. WER results of SI Hybrid systems on 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

3.2. Aurora 4 experiments

The excitation features were investigated in Aurora 4 task as well. Aurora 4 is a noise robust continuous speech recognition task, the size of vocabulary is 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 by Aurora 4 task: 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 are the same utterances but divided into two parts: one part was from the primary Sennheiser microphone and the other was from a secondary microphone which includes the convolutional distortion. 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 two channels and each channel includes clean condition plus 6 noise conditions which are same as the training data, thus includes 14 test set in total.

In this work the multi-condition training set was used for system training. The vocal tract based feature used in the model training is 25-dimensional FBANK. The process of the vocal tract based features is same as the one proposed in [31]. 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. In this work, the DNN hybrid system with SAT was used to train the Aurora 4 acoustic model. To each speaker and noise condition, a Global CMLLR transforms was trained and cascaded with the LDA+STC transforms to normalized the speaker and noise environment factors. This transformed feature vector was concatenated with the excitation feature vector proposed by this work as the input of 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 DBN based pre-training was used to initialize the DNN. Both cross-entropy based training and the sMBR based sequence training were used for fine-tune. The results are given in table 2.

Table 2 indicates that excitation features significantly reduced the WER of Aurora 4 task, which is consistent with the AMI meeting transcription results.

Table 2. WER(%) 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

4. CONCLUSION

This work introduces the robust excitation based features as the complements of the traditional vocal tract based acoustic features to improve the performance of the state-of-the-art DNN based ASR system. The suggested excitation features have been investigated and they were evaluated on two very different ASR task: the AMI meeting transcription and Aurora 4. The experimental results showed that the proposed excitation features can significantly improve the performance of various ASR task. It is worth mentioning the low average WER in Aurora 4, which supports our argument that the extraction of the excitation features was indeed robust.

5. REFERENCES

[1] 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.

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

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

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

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

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

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

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

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

[19] 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.

[20] 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.

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

[22] Steve Renals, Thomas Hain, and Herve Bourlard, "Recognition and understanding of meetings the AMI and AMIDA projects," in *ASRU, IEEE Workshop on*. IEEE, 2007, pp. 238–247.

[23] Catherine Breslin, KK Chin, Mark J F Gales, and Kate Knill, "Integrated online speaker clustering and adaptation.," in *Proc. ISCA Interspeech*, 2011, pp. 1085–1088.

[24] Thomas Hain, Lukas Burget, John Dines, Philip N Garner, Frantisek Grezl, Asmaa El Hannani, Marijn Huijbregts, Martin Karafiat, Mike Lincoln, and Vincent Wan, "Transcribing meetings with the AMIDA systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 486–498, 2012.

[25] Xie Chen, Mark Gales, Kate Knill, Catherine Breslin, Langzhou Chen, K.K. Chin, and Vincent Wan, "An initial investigation of long-term adaptation for meeting transcription," in *Proc. INTERSPEECH*, 2014.

[26] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al., "The ICSI meeting corpus," in *Proc. ICASSP*. IEEE, 2003, vol. 1, pp. I–364.

[27] John S Garofolo, Christophe Laprun, Martial Michel, Vincent M Stanford, and Elham Tabassi, "The NIST meeting room pilot corpus.," in *LREC*, 2004.

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

[29] Thomas Hain, Vincent Wan, Lukas Burget, Martin Karafiat, John Dines, Jithendra Vepa, Giulia Garau, and Mike Lincoln, "The AMI system for the transcription of speech in meetings," in *Proc. ICASSP*. IEEE, 2007, vol. 4, pp. IV–357.

[30] 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.

[31] Shakti P. Rath, Daniel Povey, Karel Vesely, and Jan Honza Cernocky, "Improved feature processing for dep neural networks," in *Proc. INTERSPEECH*, 2013.