Smoothed Nonlinear Energy Operator-Based Amplitude Modulation Features for Robust Speech Recognition

Md. Jahangir Alam[1,2], Patrick Kenny[2], and Douglas O'Shaughnessy[1]

[1] INRS-EMT, University of Quebec, Montreal (QC) Canada
[2] CRIM, Montreal (QC) Canada
{jahangir.alam,patrick.kenny}@crim.ca, dougo@emt.inrs.ca

Abstract. In this paper we present a robust feature extractor that includes the use of a smoothed nonlinear energy operator (SNEO)-based amplitude modulation features for a large vocabulary continuous speech recognition (LVCSR) task. SNEO estimates the energy required to produce the AM-FM signal, and then the estimated energy is separated into its amplitude and frequency components using an energy separation algorithm (ESA). Similar to the PNCC (Power Normalized Cepstral Coefficients) front-end, a medium duration power bias subtraction (MDPBS) is used to enhance the AM power spectrum. The performance of the proposed feature extractor is evaluated, in the context of speech recognition, on the AURORA-4 corpus, which represents additive noise and channel mismatch conditions. The ETSI advanced front-end (ETSI-AFE), power normalized cepstral coefficients (PNCC), Cochlear filterbank cepstral coefficients (CFCC) and conventional MFCC and PLP features are used for comparison purposes. Experimental speech recognition results on the AURORA-4 task depict that the proposed method is robust against both additive and different microphone channel environments.

Keywords: Amplitude modulation, SNEO, Speech recognition, AURORA-4.

1 Introduction

Traditional Mel Frequency Cepstral Coefficients (MFCCs) [1] and Perceptual Linear Prediction (PLP) [2] features are frequently used as a low-dimensional set of features to represent short segments of speech. Since it was first conceived in 1974, MFCC has remained a powerful sound representation tool as it partly mimics human perception of sound color [3], and thus is popular in the signal processing community in almost its original form. MFCCs and PLP features along with the standard Hidden Markov Model (HMM)-based speech recognizer perform well if the training and test environments are the same. Different operating conditions during signal acquisition (e.g., channel response, handset type, additive background noise, reverberation, etc.) lead to feature mismatch across training and testing and thereby degrade the performance of the MFCCs and PLP-based speech recognition systems.

The methods to compensate for the effects of environmental mismatch can be implemented at the front-end (or feature extractor) or at the back-end or both. The main

T. Drugman and T. Dutoit (Eds.): NOLISP 2013, LNAI 7911, pp. 168–175, 2013.
© Springer-Verlag Berlin Heidelberg 2013

goal of a robust feature extractor for a recognition task is to develop features that retain useful variability in speech while minimizing variability due to environmental mismatch. Various robust feature extractors are employed in speech recognition tasks such as the ETSI advanced front-end (ETSI-AFE) [4], power normalized cepstral coefficients (PNCC) [5], and the robust feature extractors proposed in [6, 7, 8].

Amplitude modulation-frequency modulation (AM-FM) of speech signal plays an important role in speech perception and recognition [8]. The AM-FM model has been successfully used in various areas of signal processing. Specifically in speech processing this model has been applied for speech analysis and modeling [9, 10, 11, 17, 19], speech synthesis [10], emotion, speech and speaker recognition [12-13, 14-15, 16-18]. A standard approach to the AM-FM demodulation problem is to use the Hilbert transform and the related Gabor's analytic signal [20]. An alternative approach is to use a nonlinear energy operator (NEO) to track the energy required to generate an AM-FM signal and separate it into amplitude and frequency components. The NEO approach to demodulation has many attractive features such as simplicity, efficiency, and adaptability to instantaneous signal variations [9]. In this paper we use smoothed nonlinear energy operator (SNEO) [20, 21]-based amplitude modulation features for a robust large-vocabulary continuous speech recognition (LVCSR) task. The advantage of SNEO (or NEO) is that it uses only a few samples of the input signal to estimate the energy required to generate an AM-FM signal and separate it into amplitude and frequency components without imposing any stationarity assumption as done by linear prediction or Fourier transforms [8]. The SNEO approach has smaller computational complexity and faster adaptation due to its instantaneous nature [28]. Since SNEO (or NEO) uses only a few samples to estimate the energy, it is sensitive to noise. We use a medium duration power bias subtraction (MDPBS) technique, proposed in [5], to enhance estimated AM power. Power function nonlinearity with a coefficient of 0.07 is applied as it has been found in [5] that it is more robust than the logarithmic nonlinearity used in a conventional MFCC framework. The final features are obtained by taking the Discrete Cosine Transform (DCT) and normalizing the features using the full utterance-based cepstral mean normalization method.

The AURORA-4 LVCSR corpus [22] is used for performance evaluation of the proposed feature extractor. To compare the performances, the following front-ends are used: conventional MFCC, PLP, ETSI-AFE [4], power normalized cepstral coefficient (PNCC) [5], Cochlear filterbank cepstral coefficients (CFCC) [24], and the robust front-end (RFE) of [6]. Experimental results on the AURORA-4 LVCSR task show that the proposed feature extractor outperforms all the front-ends mentioned above.

2 Overview of the Proposed Feature Extractor

The various steps of the proposed feature extractor are shown in Fig. 1. In this method, processing of a speech signal begins with pre-processing (including DC removal and pre-emphasis, typically using a first-order high-pass filter with a transfer function of $(1-0.97z^{-1})$). The pre-processed speech signal is then framed (analysis frame length is 25 msec with a frame shift of 10 msec) and windowed using a Symmetric Hamming window. Each frame of the speech signal is then decomposed into a

C-channel (here, C = 40 is used) gammatone filterbank covering the frequency range of 100-3800 Hz (sampling frequency = 8000 Hz). The AM power spectrum for each gammatone channel is then estimated using the smoothed nonlinear energy operator (SNEO). Before applying the medium duration power bias subtraction (MDPBS) [5] to enhance the AM power spectrum, the AM power across each frame and channel is normalized using 95th percentile power [5]. The 13-dimensional static features, obtained after applying a power function nonlinearity, using a coefficient of 0.07 and discrete cosine transform (DCT), are normalized using the conventional cepstral mean normalization method.

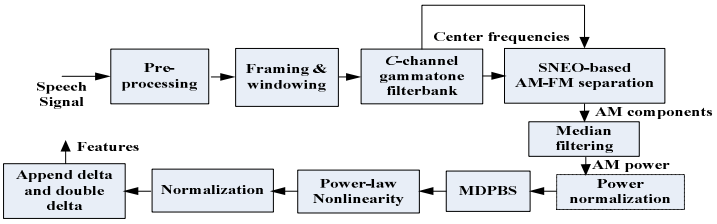

Fig. 1. Block diagram showing various steps of the proposed robust feature extractor

2.1 SNEO-Based AM-FM Separation

Extensive research by Teager resulted in a nonlinear approach for computing the energy of a signal denoted as the nonlinear energy operator (NEO) or Teager Kaiser Energy operator (TKEO) [23]. The NEO uses only a few samples of the input signal to track the energy required to generate an AM-FM signal and separate it into amplitude and frequency components in a nonlinear manner, which provides an advantage over the conventional Fourier transform (FT) or linear prediction (LP) methods in capturing the energy fluctuations. Let $x(c,n)$ represent the speech frame of the cth channel, where $c = 1,2,...,C$ is the channel (or filterbank) index of the C-channel gammatone filterbank, $n = 1,2,...,N$ is the discrete time index, N is the frame length in samples and C is the number of channels of the gammatone filterbank. Standard NEO (or TKEO) of $x(c,n)$ can be expressed as a special case of the following kth order ($k=0,1,2,...$) and lth lag ($l=1,2,3,...$) generalized discrete energy operator:

$$\Psi_{k,l}\left(x(c,n)\right) = x(c,n)\,x(c,n+k) - x(c,n-l)\,x(c,n+k+l). \tag{1}$$

For k=0 and l=1, eqn. (1) reduces to the standard NEO or TKEO. The NEO has the problem of cross terms and few negative values. To alleviate these problems we use the smoothed NEO (SNEO) [20, 21], which is expressed as:

$$\Psi^{s}_{0,1}\left(x(c,n)\right) = \Psi_{0,1}\left(x(c,n)\right) \otimes w(n), \tag{2}$$

where $w(n)$is the smoothing window and $\otimes$ represents the convolution operator. For smoothing, a Bartlett window was used in [21], whereas in [20] a 7-point binomial smoothing filter with impulse response (1, 6, 15, 20, 15, 6, 1) was applied. In this work we use the latter smoothing filter. Since SNEO (or NEO) is an energy operator and energy is a positive quantity, in order to avoid any negative values in eqn. (1) (if $x(c,n)x(c,n+k) < x(c,n-l)x(c,n+k+l)$ for k=0, l=1) we have taken the absolute values of eqn. (1) [25, 8]. Now, for the cth channel, the AM and FM components can be estimated using the discrete energy separation algorithm (DESA) when k=0, l=1as follows [20]:

$$\left|\hat{a}(c,n)\right| = \sqrt{\left(\frac{\Psi^{s}_{0,1}\left(x(c,n)\right)}{1-\left(1-\frac{\Psi^{s}_{0,1}\left(y(c,n)\right)+\Psi^{s}_{0,1}\left(y(c,n+1)\right)}{4\Psi^{s}_{0,1}\left(x(c,n)\right)}\right)^{2}}\right)}, \tag{3}$$

$$\left|\hat{\varphi}(c,n)\right| = \cos^{-1}\left(1-\left(\frac{\Psi^{s}_{0,1}\left(y(c,n)\right)+\Psi^{s}_{0,1}\left(y(c,n+1)\right)}{4\Psi^{s}_{0,1}\left(x(c,n)\right)}\right)\right), \tag{4}$$

where $y(c,n)) = x(c,n)) - x(c,n-1))$. In order to reduce the dynamic range, estimated AM and FM components are smoothed using a median filter with a window size of 5. For the mth speech frame the AM power for the cth channel is computed as:

$$P(m,c) = \sum_{n=1}^{N}\left(\left|\hat{a}(c,n)\right|^{2}\right). \tag{5}$$

2.2 Normalization and Enhancement of the AM Power Spectrum

Since SNEO (or NEO) uses only a few samples to estimate the energy, it is sensitive to noise. Therefore, the AM power estimated using (5) may be corrupted due to noise. In order to compensate for the noise, a medium duration power bias subtraction (MDPBS) [5] is applied on the AM power $P(m,c)$ after normalizing by the 95th percentile power across all frames and channels [5, 8].

2.3 Post-processing

13-dimensional static features, obtained after applying a power function nonlinearity using a coefficient of 0.07 and discrete cosine transform (DCT) on the bias subtracted AM power, are normalized using the conventional cepstral mean normalization method over the entire utterance. Delta and double-delta features are computed with a 5-frame window using regression formula [27].

3 Performance Evaluation

3.1 Speech Corpus and Experimental Setup

The AURORA-4 continuous speech recognition corpus is derived from the Wall Street Journal (WSJ0) corpus. 14 evaluation sets were defined in order to study the degradations in speech recognition performance due to microphone conditions, filtering and noisy environments [22]. The 14 test sets are grouped into the following 4 families [22, 26]: Test sets A, B, C and D. For the large-vocabulary continuous speech recognition task on the AURORA-4 corpus, all experiments employed state-tied crossword speaker-independent triphone acoustic models with 4 Gaussian mixtures per state. A single-pass Viterbi beam search-based decoder was used along with a standard 5K lexicon and bigram language model with a prune width of 250 [22, 26]. The HTK (Hidden Markov Model Toolkit) recognizer [27] is employed for the recognition task.

3.2 Results and Discussion

In order to verify the effectiveness of the proposed robust feature extractor, speech recognition experiments were conducted on the AURORA-4 large vocabulary continuous speech recognition (LVCSR) corpus. Percentage word accuracy was used as a performance evaluation measure for comparing the recognition performances of the proposed method to that of the following feature extractors: MFCC, PLP, ETSI-AFE [4], power normalized cepstral coefficient (PNCC) [5], Cochlear filterbank cepstral coefficients (CFCC) [24], and the robust front-end (RFE) of [6]. Features in the MFCC and PLP front-ends are normalized using the mean and variance normalization method. CFCC and the front-end proposed in [6] utilize a short-time mean and scale normalization technique [28] to normalize the features. PNCC and the proposed method use cepstral mean normalization whereas ETSI-AFE uses a blind equalization technique, which is based on the comparison to a flat spectrum and the application of the LMS algorithm, for improving robustness of ASR systems against additive noise distortions and channel effects. Speech recognition experiments were conducted on the four test sets (A, B, C, and D) of the AURORA-4 corpus. Test set A represents the matched training/test condition (same channel) where acoustic models were trained using clean training features and recognition were performed on the clean test features. Test set B represents the mismatched training/test condition (same channel) where mismatch was created randomly adding each of the 6 noise types (car, babble, restaurant, street traffic, airport, and train-station noises) at a randomly chosen SNR between 5 and 15 dB to the test data. Training data is the same as the training data of test set A. Test set C represents the mismatched training/test condition due to different channels where acoustic models were trained using clean training features extracted from the clean training data recorded with a Sennheiser microphone and recognition was performed on the clean test features extracted from the clean test data recorded with a secondary microphone. Test set D represents the mismatched training/test condition due to additive noise and different microphone channels where acoustic models

were trained using clean training features extracted from the clean training data recorded with a Sennheiser microphone and recognition was performed on the noisy test features extracted from the noisy test data recorded with a secondary microphone.

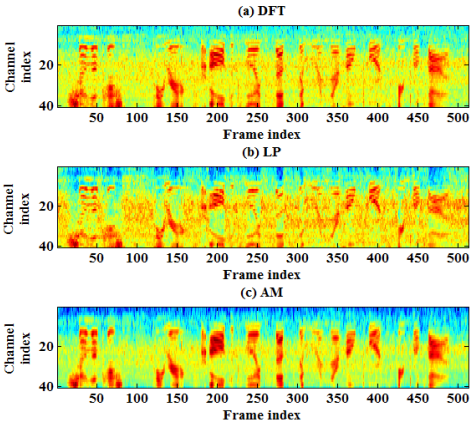

Fig. 2. Speech spectrograms after auditory filterbank integration, street noise, SNR = 5 dB, (a) DFT-based periodogram with Mel-filterbank, (b) LP spectrum with Mel-filterbank, and (c) AM power spectrum with gammatone filterbank

Table 1. Word accuracies (%) obtained by the various feature extractors on the AURORA-4 corpus. The higher the word accuracy the better is the performance of the feature extractor.

	Word Accuracy (%)				
	A	**B**	**C**	**D**	**Average**
MFCC	**90.02**	49.19	71.12	35.44	61.44
PLP(HTK)	89.72	50.41	74.44	39.64	63.55
CFCC	86.34	63.05	78.60	54.70	70.67
ETSI-AFE	88.59	69.58	79.52	61.51	74.80
PNCC [5]	88.64	69.85	81.07	60.00	74.89
RFE [6]	88.90	68.87	80.94	59.25	74.49
Proposed	87.41	**71.46**	**82.10**	**62.99**	**75.99**

Fig. 2 presents the speech auditory spectrograms of a noisy speech signal, corrupted with the street noise (SNR = 5 dB), obtained by the DFT-based periodogram, LP (linear prediction) spectrum, and AM power spectrum estimators. It is observed from this figure that compared to the other estimators, AM spectrum estimator results in a reduction of the noise while preserving the formant structure. Experimental results presented in Table 1 show that in matched environments the proposed method provides less word recognition accuracy compared to the other front-ends. It is observed from Table 1 that under mismatched environments (due to additive noise and different microphone channels) the proposed feature extractor outperformed the other feature extractors in terms of the recognition word accuracy. Therefore, the proposed method is found to be robust under environmental mismatch conditions.

4 Conclusion

A robust feature extractor that incorporates smoothed nonlinear energy operator-based amplitude modulation features for robust speech recognition is presented. Speech recognition results were reported on the AURORA-4 LVCSR corpus and performances were compared with the ETSI-AFE, PNCC, CFCC and the robust feature extractor of [6]. Experimental results depict that under mismatched condition (e.g., in test sets B, C, and D) the proposed method outperformed all the other feature extractors considered in this work in terms of the percentage word accuracy. Our future work will be to incorporate AM features presented in this paper to the feature extraction framework of [6].

References

1. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoustics, Speech, and Signal Processing 28(4), 357–366 (1980)
2. Hermansky, H.: Perceptual linear prediction analysis of speech, J. Acoust. Soc. Am. 87(4), 1738–1752 (1990)
3. Terasawa, H.: A Hybrid Model for Timbre Perception: Quantitative Representations of Sound Color and Density. Ph.D. Thesis, Stanford University, Stanford, CA (2009)
4. ETSI ES 202 050, Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; advanced front-end feature extraction algorithm; Compression algorithms (2003)
5. Kim, C., Stern, R.M.: Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring. In: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 4574–4577 (March 2010)
6. Alam, M.J., Kenny, P., O'Shaughnessy, D.: Robust Feature Extraction for Speech Recognition by Enhancing Auditory Spectrum. In: Proc. INTERSPEECH, Portland Oregon (September 2012)
7. van Hout, J., Alwan, A.: A novel approach to soft-mask estimation and log-spectral enhancement for robust speech recognition. In: Proc. of ICASSP, pp. 4105–4108 (2012)
8. Vikramjit Mitra, H., Franco, M., Graciarena, A.: Mandal, Normalized Amplitude modulation features for large vocabulary noise-robust speech recognition. In: Proc. of ICASSP, pp. 4117–4120 (2012)
9. Maragos, Kaiser, J.F., Quatieri, T.F.: On amplitude and frequency demodulation using energy operators. IEEE Trans. Signal Processing 41(4), 1532–1550 (1993)
10. Potamianos, A., Maragos, P.: Speech analysis and synthesis using an AM–FM modulation model. Speech Communication 28, 195–209 (1999)
11. Dimitriadis, D., Maragos, P.: Continuous energy demodulation methods and application to speech analysis. Speech Communication 48(7), 819–837 (2006)
12. Zhou, G., Hansen, J.H.L., Kaiser, J.F.: Nonlinear feature based classification of speech under stress. IEEE Transactions on Speech and Audio Processing 9, 201–216 (2001)
13. Gao, H., Chen, S.G.: Emotion classification of mandarin speech based on TEO nonlinear features. Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 394–398 (2007)

14. Jabloun, F., Cetin, A.E., Erzin, E.: Teager energy based feature parameters for speech recognition in car noise. IEEE Signal Processing Letters 6(10), 259–261 (1999)
15. Dimitriadis, D., Maragos, P., Potamianos, A.: Robust AM–FM features for speech recognition. IEEE Signal Processing Letters 12(9), 621–624 (2005)
16. Jankowski Jr., C.R., Quatieri, T.F., Reynolds, D.A.: Measuring fine structure in speech: Application to speaker identification. In: ICASSP 1995, Detroit, USA (May 1995)
17. Plumpe, M.D., Quatieri, T.F., Reynolds, D.A.: Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Trans. Speech and Audio Processing 7(5), 569–586 (1999)
18. Grimaldi, M., Cummins, F.: Speaker identification using instantaneous frequencies. IEEE Trans. Audio, Speech and Language Processing 16(6), 1097–1111 (2008)
19. Tsiakoulis, P., Potamianos, A.: Statistical Analysis of Amplitude Modulation in Speech Signals using an AM-FM Model. In: Proc. Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2009), Taipei, Taiwan (April 2009)
20. Potamianos, A., Maragos, P.: A comparison of energy operator and Hilbert transform approach to signal and speech demodulation. Signal Process 37(1), 95–120 (1994)
21. Mukhopadhyay, S., Ray, G.C.: A new interpretation of nonlinear energy operator and its efficacy in spike detection. IEEE Tans. on Biomedical Engg. 45(2), 180–187 (1998)
22. Parihar, N., Picone, J., Pearce, D., Hirsch, H.G.: Performance analysis of the Aurora large vocabulary baseline system. In: Proceedings of the European Signal Processing Conference, Vienna, Austria (2004)
23. Kaiser, J.F.: On a Simple Algorithm to Calculate the 'Energy' of a Signal,". In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, NM, pp. 381–384 (April 1990)
24. Li, Q(P.), Huang, Y.: Robust speaker identification using an auditory-based feature. In: Proc. ICASSP, pp. 4514–4517 (2010)
25. Kvedalen, E.: Signal processing using the Teager energy operator and other nonlinear operators, Cand. Scient Thesis, University of Oslo (May 2003)
26. Au Yeung, S.-K., Siu, M.-H.: Improved performance of Aurora-4 using HTK and unsupervised MLLR adaptation. In: Proceedings of the Int. Conference on Spoken Language Processing, Jeju, Korea (2004)
27. Young, S.J., et al.: HTK Book, Entropic Cambridge Research Laboratory Ltd., 3.4 edition (2006), http://htk.eng.cam.ac.uk/
28. Alam, M.J., Ouellet, P., Kenny, P., O'Shaughnessy, D.: Comparative Evaluation of Feature Normalization Techniques for Speaker Verification. In: Travieso-González, C.M., Alonso-Hernández, J.B. (eds.) NOLISP 2011. LNCS, vol. 7015, pp. 246–253. Springer, Heidelberg (2011)