Third-Order Moments of Filtered Speech Signals for Robust Speech Recognition

Kevin M. Indrebo, Richard J. Povinelli, and Michael T. Johnson

Dept. of Electrical and Computer Engineering, Marquette University,
Milwaukee, Wisconsin, USA
{kevin.indrebo, richard.povinelli,
mike.johnson}@Marquette.edu

Abstract. Novel speech features calculated from third-order statistics of subband-filtered speech signals are introduced and studied for robust speech recognition. These features have the potential to capture nonlinear information not represented by cepstral coefficients. Also, because the features presented in this paper are based on the third-order moments, they may be more immune to Gaussian noise than cepstrals, as Gaussian distributions have zero third-order moments. Experiments on the AURORA2 database studying these features in combination with Mel-frequency cepstral coefficients (MFCC's) are presented, and some improvement over the MFCC-only baseline is shown when clean speech is used for training, though the same improvement is not seen when multi-condition training data is used.

1 Introduction

Spectral-based acoustic features have been the standard in speech recognition for many years, even though they are based on limiting assumptions of the linearity of the speech production mechanism [1]. Specifically, mel-frequency cepstral coefficients (MFCC), which are calculated using a discrete cosine transform on the smoothed power spectrum, and perceptual linear prediction (PLP) cepstral coefficients, similar to MFCCs, but based on human auditory models, are used in almost all state-of-the-art speech recognition systems [1]. While these feature sets do an excellent job of capturing linear information of speech signals, they do not encapsulate information about nonlinear or higher-order statistical characteristics of the signals, which have been shown to exist, and are not insignificant [2-4].

As successful as MFCCs have been in the field of speech recognition, performance of state-of-the-art systems remains unacceptable for many real applications. One of the largest failings of popular spectral features is their poor robustness in the face of ambient noise. Many environments in which automatic speech recognition applications would be ideal have large amounts of background additive noise that makes voice-activated systems infeasible. In this paper, we introduce acoustic features based on higher-order statistics of speech signals. It is shown that these features, when combined with MFCC's, can produce higher recognition accuracies in some noise conditions.

The rest of the paper is as follows. Section 2 gives some background on robust speech recognition and nonlinear speech recognition. In section 3, computation of the

M. Faundez-Zanuy et al. (Eds.): NOLISP 2005, LNAI 3817, pp. 277 – 283, 2005.

proposed features is detailed. Experiments comparing the feature sets including the third-order moment features and MFCC's are presented in section 4, and are followed by the conclusion in section 5.

2 Background

2.1 Robust Speech Recognition

Robust speech recognition research has focused on subjects such as perceptually motivated features, signal enhancement, feature compensation in noise, and model adaptation. Perceptual-based features include PLP cepstral coefficients [5] and perceptual harmonic cepstral coefficients (PHCC) [6], which have been shown to be more robust than MFCCs in the presence of additive noise. Signal enhancement and feature compensation include techniques like spectral subtraction [7] and iterative wiener filtering [8], as well as more advanced algorithms such as SPLICE (stereo-based piecewise linear compensation in environments) [9]. While these techniques focus on adapting the extracted features, model adaptation methods such as MLLR and MAP [10] attempt to adjust the model parameters to better fit the noisy signals.

Though some progress has been made, the performance of speech recognition systems in noisy environments is still far from acceptable. Word error rates for a standard large vocabulary continuous speech recognition (LVCSR) task like recognition of the 5,000 word Wall Street Journal corpus can drop from under 5% to over 20% when Gaussian white noise is added at a signal-to-noise-ratio (SNR) of +5dB, even with compensation techniques [9]. Even continuous digit recognition word error rates often exceed 10% when faced with high noise levels [11].

2.2 Nonlinear Features for Speech Recognition

Recently, work has been done to investigate the efficacy of various feature sets based on nonlinear analysis. Dynamical invariants based on chaos theory [12], such as Lyapunov exponents and fractal dimension have been used to augment the standard linear feature sets [13], as well as nonlinear polynomial prediction coefficients [14]. In [15], an AM-FM model of speech production is exploited for extraction of nonlinear features. Also, Phase space reconstruction has been used for statistical modeling and classification of speech waveforms [16].

In [17], reconstructed phase spaces built from speech signals that have been subband filtered were used for isolated phoneme classification, showing improved recognition accuracies over fullband signal phase space reconstruction features. However, this approach is infeasible for continuous speech recognition because of its high computational complexity. In this paper, nonlinear features from subbanded speech signals that are much simpler to compute are introduced.

3 Third-Order Moment Feature Computation

An approach based on time-domain filtering of speech signals is taken for computation of the nonlinear features. An utterance is parameterized by first filtering the signal into P subbands that have cutoffs and bandwidths derived from the Mel-scale. Each of these signals is then broken into frames with lengths of 25.6 ms, updated every 10 ms. The third-order moment of the signal amplitudes of each of these channels is calculated for each frame, and the set of these coefficients form a feature vector. Log energy of the unfiltered signal frame is appended to these features, which are then orthogonalized using a principle component analysis (PCA). In the experiments presented in this paper, 20 filter channels are used, and the PCA reduces the dimension third-order moment feature space to 13.

Because much of the information that distinguishes speech sounds is contained in the power spectrum, it is not expected that these features by themselves would carry enough information to compete with MFCC features. Therefore, the proposed features are appended to the baseline MFCC feature vector for modeling and recognition of speech.

There are two advantages to this approach. First, nonlinear information that may be useful for recognition that is not captured by traditional features is added to the recognizer. Also, because some types of noise have approximately Gaussian statistical distributions, and Gaussian distributions have zero third-order moments, the proposed features my be less affected by additive noise than MFCC's. This conjecture is tested by comparing a combined feature set of MFCC's and the proposed features to a baseline feature set of only MFCC's for use in noisy speech recognition.

4 Experiments

The preliminary recognition experiments are run using the AURORA2 database [18]. This corpus contains utterances of connected digits corrupted with different types and levels of noise. There are eleven words: the digits zero through nine and "oh". Two sets of experiments were run. In the first set the models were trained using clean speech signals, and tested on test set A, which contains four different types of noise

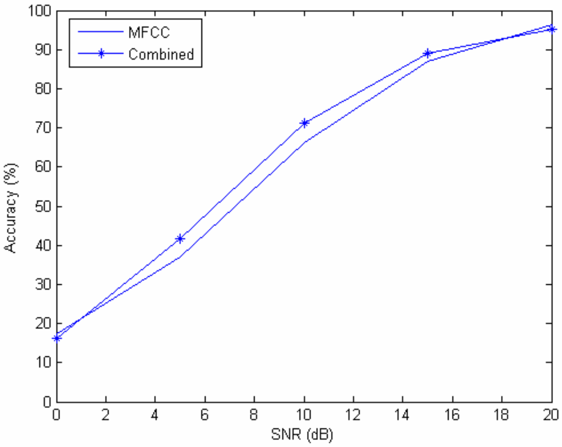

Fig. 1. Recognition accuracies for speech corrupted by subway noise

at varying SNR levels. The second set of experiments used models trained on the multi-condition training set in AURORA2, and the tests were performed on test set A and test set B, which has four different types of noise. The multi-condition training set has the same noise types as test set A, providing a matched noisy training-test scenario. The noise types in test set B are not included in any training signals. HTK [19] is the software used for experimentation. Each word is modeled using a 16-state left-to-right diagonal covariance Hidden Markov Model (HMM). Additionally, a 3-state silence model and single-state short pause model are implemented. The frame rate is 10 ms, with frame lengths of 25.6 ms.

Two types of feature sets are used. The baseline feature vector is a 39-element vector of 12 MFCC's, log energy, and the first and second time derivatives. The second feature

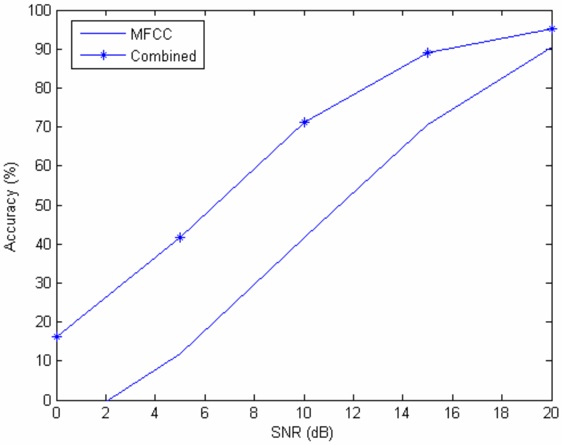

Fig. 2. Recognition accuracies for speech corrupted by babble noise

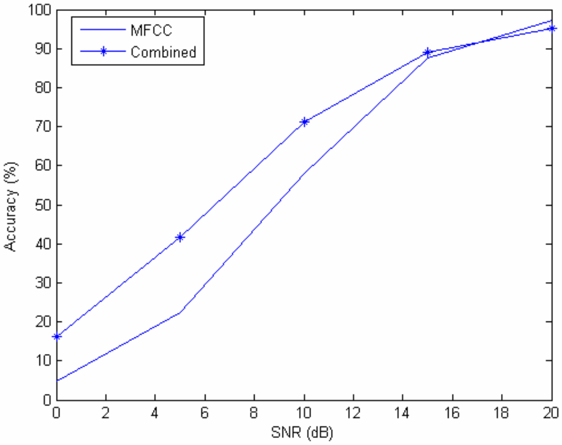

Fig. 3. Recognition accuracies for speech corrupted by car noise

set is a 45-element vector composed of the 39-coefficient MFCC vector concatenated with 6 coefficients from the PCA of the third-order moment space.

Figures 1-4 show the recognition accuracies for the two feature types on the four different noise types of AURORA's test set A, using models trained on clean speech signals. These noises are subway, babble, car noise, and exhibition hall, respectively. The accuracies are plotted against the SNR levels, ranging from 0 to 20 dB. It can be seen that, except for the babble noise case, the MFCC-only features give better recognition accuracies at 20 dB SNR. The MFCC and third-order moment concatenation feature vector, however, outperforms the MFCC-only set in most of the lower SNR cases.

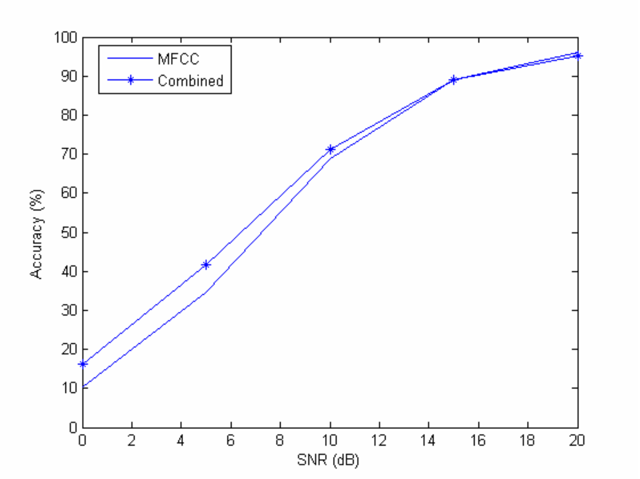

Fig. 4. Recognition accuracies for speech corrupted by exhibition hall noise

Table 1. Average recognition accuracies for models trained on corrupted speech

Feature type	Test set A	Test set B
MFCC's	88.22%	84.10%
Combined features	80.98%	61.37%

Table 1 shows the accuracies of models trained on the multi-condition training set and tested on both test sets A and B for the MFCC feature set and the combined feature set, averaged over all the noise types and SNR levels from 0 to 20 dB. This table shows that when the models are trained on speech corrupted with different types and levels of noise, the addition of the third-order moment features does not improve upon the MFCC baseline, even degrading the performance significantly.

5 Conclusion

A new type of acoustic feature extraction method was presented based on higher-order statistics of subband filtered speech signals, and tested on noisy speech signals. The results show that the combination of traditional MFCC features and these new features can improve the robustness of speech recognition systems when the speech models are trained on clean speech data. The largest improvement is seen when the speech signals used for recognition are corrupted by babble noise. However, when the speech models are trained on clean speech, the performance of the recognition degrades compared to MFCC only features. For these features to be useful in real systems, some adaptive combination may be necessary, so that information from third-order moment features is only used when it will improve the recognition estimates of the recognition system.

References

[1] B. Gold and N. Morgan, *Speech and Audio Signal Processing*. New York, New York: John Wiley and Sons, 2000.

[2] M. Banbrook and S. McLaughlin, "Is Speech Chaotic?," presented at IEE Colloquium on Exploiting Chaos in Signal Processing, 1994.

[3] M. Banbrook, S. McLaughlin, and I. Mann, "Speech characterization and synthesis by nonlinear methods," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 1 -17, 1999.

[4] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," presented at NATO ASI on Speech Production and Speech Modelling, 1990.

[5] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech recognition," presented at Journal of the Acoustical Society of America, 1990.

[6] L. Gu and K. Rose, "Perceptual harmonic cepstral coefficients for speech recognition in noisy environments," presented at IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01), Salt Lake City, UT, 2001.

[7] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113-120, 1979.

[8] K. Yu, B. Xu, M. Dai, and C. Yu, "Suppressing cocktail party noise for speech recognition," presented at 5th International conference on signal processing (WCCC-ICSP 2000), Beijing, China, 2000.

[9] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-Vocabulary Speech Recognition Under Adverse Acoustic Environments," presented at Internation Conference on Spoken Language Processing (ICSLP), Beijing, China, 2000.

[10] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book," 1997.

[11] C. Meyer and G. Rose, "Improved Noise Robustness By Corrective and Rival Training," presented at International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03), 2003.

[12] E. Ott, *Chaos in dynamical systems*. Cambridge, England: Cambridge University Press, 1993.

[13] V. Pitsikalis and P. Maragos, "Speech analysis and feature extraction using chaotic models," presented at International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2002.

[14] X. Liu, R. J. Povinelli, and M. T. Johnson, "Vowel Classification by Global Dynamic Modeling," presented at ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP), Le Croisic, France, 2003.

[15] D. Dimitriadis, P. Maragos, and A. Potamianos, "Modulation features for speech recognition," presented at International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2002.

[16] M. T. Johnson, R. J. Povinelli, A. C. Lindgren, J. Ye, X. Liu, and K. M. Indrebo, "Time-Domain Isolated Phoneme Classification using Reconstructed Phase Spaces," *IEEE Transactions on Speech and Audio Processing*, in press.

[17] K. M. Indrebo, R. J. Povinelli, and M. T. Johnson, "Sub-banded Reconstructed Phase Spaces for Speech Recognition," *Speech Communication*, in press.

[18] D. Pearce and H. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," Beijing, China, 2000.

[19] "HTK Version 2.1," Entropic Cambridge Research Laboratory Ltd., 1997.