

SPECTRAL MAXIMA REPRESENTATION FOR ROBUST AUTOMATIC SPEECH RECOGNITION

J.Sujatha* , K.R.Prasanna Kumar[†] , K.R.Ramakrishnan[‡] and N.Balakrishnan[§]

Indian Institute of Science, Bangalore, INDIA 560 012.

Abstract

In the context of automatic speech recognition, the popular Mel Frequency Cepstral Coefficients(MFCC) as features, though perform very well under clean and matched environments, are observed to fail in mismatched conditions. The spectral maxima are often observed to preserve their locations and energies under noisy environments, but are not presented explicitly by the MFCC features. This paper presents a framework for representing the maxima information for robust recognition in the presence of additive White Gaussian Noise(WGN). For the task of phoneme based Isolated Word Recognition (IWR) under different Signal to Noise Ratio (SNR) environments, the results show an improved recognition performance. The cepstral features are computed from a reconstructed spectrogram by fitting gaussians around the spectral maxima. In view of the inherent robustness and easy trackability of the maxima, this opens up interesting avenues towards a robust feature representation as well as preprocessing techniques.

1. Introduction

Robustness to additive environmental noise is a prime requirement for practical applications of speech recognition. One of the points highlighted by a survey of research techniques for noisy speech recognition is that, it is essential to give more importance to high Signal to Noise Ratio (SNR) portions of speech in decision making [1]. Dominant spectral peaks are such high SNR values in the original spectrum. Effect of additive noise on the spectrogram of a speech signal shows that the spectral valleys are more corrupted than the spectral peaks. The noise tends to flatten the valleys in the spectrum, thereby reducing the variance of the noise-corrupted speech. Consequently, the spectral peak-to-valley ratio is also reduced distorting the spectral contrast.

The issue of making use of spectral peak information in a speech recognition system has been previously addressed in [2] and [3]. However, applicability of spectral

maxima as feature vector directly presents some problems such as their non-uniform length across the frames, as well as identification of spurious maxima in the presence of noise. This paper proposes a framework for incorporating the maxima information into the standard Mel Frequency Cepstral Coefficient (MFCC) feature vectors. Towards achieving this, the speech spectrum is reconstructed by fitting gaussians of uniform variance around the spectral maxima and is fed as input to the mel-banks in a standard MFCC-based front-end of a speech recogniser. Improved recognition results, with both clean as well as additive White Gaussian Noise (WGN) corrupted speech, have been recorded with MFCC features computed on such reconstructed spectrograms.

A signal can be represented quite efficiently in terms of nonorthogonal gaussians [4]. Ideally, it is possible to derive the parameters of the optimal gaussian basis functions that represent the signal most efficiently, i.e. minimum error for a given number of basis functions. Significant studies have been made towards obtaining optimal algorithms for finding the best fit in terms of gaussians, in the least square sense. However, this presents a difficult nonlinear optimization problem that is computationally prohibitive. Moreover, in the speech recognition front-end, the gross spectral characteristics of the speech signal are presented through a bank of overlapping, mel-scale spaced, filters to the Hidden Markov Models (HMM), after decorrelating them with the Discrete Cosine Transform (DCT). A complete or exact reconstruction of the spectrum is hence not essential for speech recognition. Such gross information about the spectral characteristics of the given speech signal is reliably conveyed by the spectral maxima. Also, the spectral maxima deserve a better fit since they are perceived with greater accuracy [5], which in this case, is ensured by the least error at the spectral peaks. Though computation of maxima brings in an overhead in terms of book-keeping, the experimental results show that it has distinct advantages in terms of robustness.

Section 2 details the computation of spectral maxima and the reconstruction procedure using gaussians. Section 3 describes the experimental set up and presents the recognition results in terms of word and phoneme recognition for the task of phoneme based Isolated Word

* Dept. of Electrical Engineering

[†] Dept. of Aerospace Engineering

[‡] Dept. of Electrical Engineering

[§] Dept. of Aerospace Engineering & SERC

Recognition (IWR), for clean and additive WGN corrupted speech at different SNR. A concluding discussion on the issues involved in further improving the robustness of speech recognition with the robust spectral maxima representation is presented in section 4.

2. Maxima computation and spectral reconstruction

2.1. Definition and computation of maxima

A function $f(x)$ has a relative (or local) maximum at x_0 if there is some interval (r, s) containing x_0 for which $f(x_0) > f(x)$ for all x between r and s for which $f(x)$ is defined. Similarly, $f(x)$ has a relative (or local) minimum at x_0 if there is an interval (r, s) containing x_0 for which $f(x_0) < f(x)$ for all x between r and s for which $f(x)$ is defined. Relative extremum means either a relative maximum or a relative minimum. By the first derivative test, relative extrema occur where $f'(x)$ changes sign. Once the extrema are found, the local maximum or minimum can be found by the second derivative test. i.e. if $f'(x_0) = 0$ and $f''(x_0) > 0$, then $f(x)$ has a local minimum at x_0 . If $f'(x_0) = 0$ and $f''(x_0) < 0$, then $f(x)$ has a local maximum at x_0 . For the given discrete sequence, the first and second order derivatives are approximated by their first and second order differences.

Spectral maxima are the maxima computed from the spectrum of the given speech signal and are known to carry significant information of the underlying speech. They are characterised by their position as well as magnitude. Techniques like 'sine wave modelling' for speech synthesis, demonstrate the reconstruction capability of the spectral peaks and hence their importance [6]. In addition to the spectral energy content, the spectral shape information is also conveyed by the spectral maxima.

2.2. Reconstruction of the spectrum with spectral maxima

Let $S(k)$ be the spectrum of the time sampled speech signal $s(n)$ as computed by the N point discrete fourier transform,

$$s(k) = \sum_{n=1}^{n=N} s(n)e^{-j2\pi k(n/N)} \quad (1)$$

The spectral maxima are those points on the spectrum satisfying,

$$S'(k) = 0 \quad (2)$$

$$S''(k) < 0 \quad (3)$$

The spectrum is approximated as a summation of the gaussians, placed at the locations of the spectral maxima and can be represented by,

$$\hat{S}(k) = \sum_{i=1}^{i=M} c_i \mathcal{N}_{\sigma_i}(k - x_i) \quad (4)$$

where M is the number of maxima in the given spectral frame, c_i is the maxima amplitude, x_i is the maxima location and σ_i is the variance of the i^{th} gaussian defined by,

$$\mathcal{N}_{\sigma_i}(x) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_i^2}} \quad (5)$$

Such a reconstruction, from the not necessarily uniformly spaced spectral peaks, appears to be a wide-band version of the original spectrogram. Though coarse in structure, it is acceptable for the task of recognition since the conventional MFCC-based front-end of a speech recogniser smoothen the speech spectrum through mel-bank filters as mentioned previously. Optimal results were obtained when the width of the gaussians was uniformly chosen in the range of 250Hz to 300Hz during reconstruction.

The original and the reconstructed spectrograms for the utterance "ONE" are shown in Fig.1 and Fig.2 respectively.

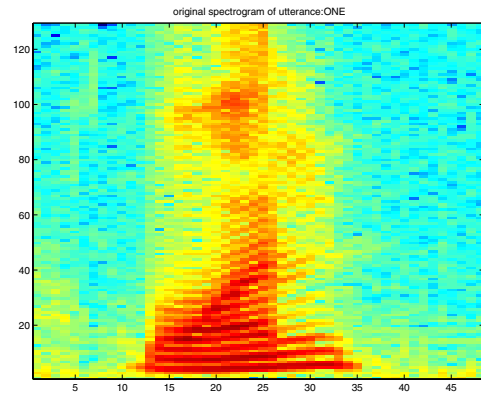


Figure 1: original spectrogram

3. Experimental setup and recognition results

Recognition experiments were conducted with MFCC features derived out of both the original and the reconstructed spectrograms. The performance of the two were compared with both clean as well as additive WGN corrupted test utterances at four different values of SNR.

3.1. Experimental setup

A phoneme-based isolated word recognizer for the recognition of ten digits(0-9) was built using the HTK toolkit [7]. Continuous density, left-to-right, HMM models were

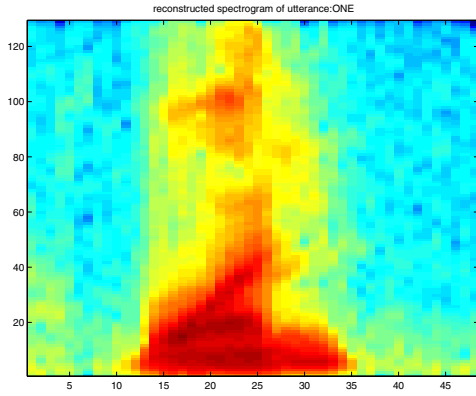


Figure 2: reconstructed spectrogram

trained for 22 monophones with a word level decoder. Each phoneme was modelled with 3 emitting states and each state with 4 gaussian mixtures.

The standard MFCC (MFCC) features were computed from the fourier magnitude spectrum of each hamming windowed frame of the speech signal. Frames were of duration 32ms with an overlap of 16ms. DCT was used to decorrelate the feature vector.

For computing the MFCC features from the reconstructed spectrum(MFCC_R), first the spectral maxima were computed from the fourier magnitude spectrum of each hamming windowed frame of the speech signal. All the available maxima in the given frame were used for the reconstruction with gaussians of width 250Hz. It was observed that on an average, 30-40 maximas could be picked in each frame. The frame duration and overlap were 32ms and 16ms as before.

There were 2240 training utterances and 2260 for testing chosen from the corresponding sets of the TIDIGITS database. The 20kHz sampled database was down sampled to 8kHz to get telephone bandwidth speech.

For experimental purposes, noisy utterances were generated by adding white gaussian noise segmentally to the clean speech utterances, to meet the SNR requirement¹.

3.2. Recognition results

This section presents the recognition results recorded with the standard(MFCC) and the reconstructed spectrogram based (MFCC_R) feature vectors. The number of static cepstral coefficients was kept to 12. Recognition scores were recorded with only the static features and then the same appended with delta(Δ) and acceleration ($\Delta\Delta$) coefficients. Experiments were also repeated with and without Cepstral Mean Subtraction (CMS) on the feature vectors. The word and phoneme recognition scores are tabu-

¹WGN source selected from an archive of additive noise sources developed at the Robust Speech Processing Laboratory, University of Colorado at Boulder(http://csrl.colorado.edu/rspl/rspl_software.html)

lated for each of the cases as shown in following tables.

Table 1: Word Recognition Accuracy without CMS

Feature Type	MFCC	MFCC_R	MFCC + Δ + $\Delta\Delta$	MFCC_R + Δ + $\Delta\Delta$
Feature Length	12	12	36	36
Test data ↓				
Clean speech	97.92	97.04	99.20	99.56
(SNR 20dB)	92.3	91.59	96.28	97.7
(SNR 10dB)	49.38	47.04	64.07	67.35
(SNR 5dB)	31.99	27.52	48.54	49.65
(SNR 0dB)	22.83	22.08	36.06	36.46

Table 2: Phoneme Recognition Accuracy without CMS

Feature Type	MFCC	MFCC_R	MFCC + Δ + $\Delta\Delta$	MFCC_R + Δ + $\Delta\Delta$
Feature Length	12	12	36	36
Test data ↓				
Clean speech	76.27	79.31	91.15	91.54
(SNR 20dB)	65.87	68.93	84.80	84.80
(SNR 10dB)	54.53	55.29	68.99	69.04
(SNR 5dB)	50.69	51.32	61.93	62.27
(SNR 0dB)	48.56	47.22	52.64	53.04

Table1 records the word recognition scores of the two feature sets in comparison. A comparable performance is observed between the MFCC and the MFCC_R features when only the static features are used. Appending the delta and acceleration coefficients however, shows an improvement in the performance of MFCC_R features against that of MFCC features, for the recognition of both clean as well as noisy speech. Corresponding phoneme recognition scores are tabulated in Table2 and show similar behaviour to that of word recognition results of Table1. More importantly, when CMS was applied to both the types of feature vectors, the MFCC_R features record a significant improvement in word recognition over the standard MFCC features as highlighted in Table3. It is also noted that the improvement in recognition is observed both with static features of length 12 as well as when appended with their first and second derivatives. Though marginal, a similar improvement in phoneme recognition scores by the MFCC_R features over the MFCC features is shown by Table4.

The respective highest word recognition scores have been recorded by the two feature sets when appended with their first and second derivatives and further by applying CMS. The corresponding word error rates of the

Table 3: Word Recognition Accuracy with CMS

Feature Type	MFCC	MFCC_R	MFCC + Δ + $\Delta\Delta$	MFCC_R + Δ + $\Delta\Delta$
Feature Length	12	12	36	36
Test data \downarrow				
Clean speech	98.67	98.10	99.12	99.6
(SNR 20dB)	95.22	94.38	96.64	97.88
(SNR 10dB)	65.49	71.86	73.14	76.19
(SNR 5dB)	49.42	55.13	57.65	60.31
(SNR 0dB)	39.87	42.26	48.23	49.07

Table 4: Phoneme Recognition Accuracy with CMS

Feature Type	MFCC	MFCC_R	MFCC + Δ + $\Delta\Delta$	MFCC_R + Δ + $\Delta\Delta$
Feature Length	12	12	36	36
Test data \downarrow				
Clean speech	83.29	83.82	91.15	91.12
(SNR 20dB)	73.51	74.6	84.85	85.52
(SNR 10dB)	58.03	61.22	72.52	72.67
(SNR 5dB)	53.13	54.96	64.9	64.89
(SNR 0dB)	50.51	50.58	58.06	58.61

two feature sets are plotted in Fig.3 for comparison which demonstrates the superiority of the MFCC_R features over MFCC features for clean as well as noisy speech recognition.

4. Conclusion

Robustness is a key issue for the practical application of automatic speech recognisers. A reconstruction of the speech spectrum based on spectral maxima, which are high SNR points, has been proposed in this paper. The reconstruction is achieved by constructing uniform width gaussians around the spectral maxima points. It is shown that this provides a suitable framework for constructing a robust frontend for a HMM based speech recogniser. Results on the task of IWR are presented and have demonstrated the equivalence in performance with clean speech and more significantly, an improved robust performance with additive WGN corrupted speech over different SNR values. It is also shown that the CMS technique improves the recognition performance to a greater extent with the proposed MFCC_R features in comparison to that of the standard MFCC features.

It is noted from the above experimental results that the spectral maxima convey inherently robust information for speech recognition. However, it is observed that

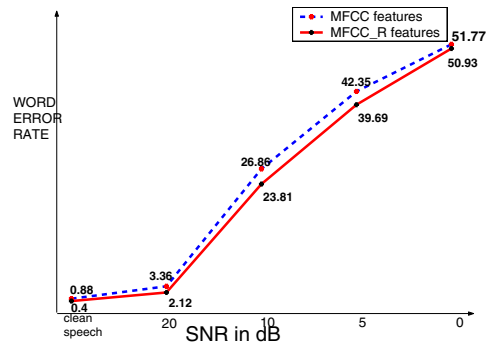


Figure 3: word error rate for MFCC & MFCC_R features with CMS

few spurious maxima are introduced in the spectrum by the additive noise. As a result, it is important to verify the authenticity of the maxima, as well as reestimate their value for better reconstruction, in the presence of noise. The original energy contours are well preserved in the spectrogram of a noisy speech signal and hence the maxima are well trackable. In this context, it is relevant and important to further explore the presented concepts for better exploitation of the potential of the spectral maxima towards robust automatic speech recognition.

5. References

- [1] Y.Gong, "Speech recognition in noisy environments: a survey," *speech communication*, vol. 16, no. 3, pp. 261–291, April 1995.
- [2] M.Padmanabhan, "Spectral peak tracking and its use in Speech Recognition", *Proceedings ICSLP*, 2000.
- [3] B.Strope and A.Alwan, "Robust word recognition using threaded spectral peaks," in *Proceedings ICASSP*, vol. 2, pp. 625–628, 1998.
- [4] J.Ben-Arie and K.R.Rao, "Nonorthogonal signal representation by gaussians and gabor functions," *IEEE Transactions on Circuits and Systems-II*, vol. 42, no. 6, pp. 402–413, June 1995.
- [5] F.J.Owens and R.Lingard, "Analytic pole-zero modelling of speech spectra," in *Proceedings ICASSP*, vol. 7, pp. 1577–1580, 1982.
- [6] J.Barker and M.P.Cooke, "Modelling the recognition of spectrally reduced speech," in *Proceedings Eurospeech*, pp. 2127–2130, 1997.
- [7] <http://htk.eng.cam.ac.uk/>.