Institut Eurécom
Department of Multimedia Communications
2229, route des Crétes
B.P. 193
06904 Sophia-Antipolis
FRANCE

Research Report RR-04-119

On Desensitizing the Mel-Cepstrum to Spurious Spectral Components for Robust Speech Recognition

Date 20th September 2004

Vivek Tyagi, Christian Wellekens

Tel : (+33) 4 93 00 26 26
Fax : (+33) 4 93 00 26 27
Email : {Vivek.Tyagi , Christian.Wellekens}@eurecom.fr

[1]Institut Eur´ecom's research is partially supported by its industrial members: Bouygues T´e´ecom,Fondation d'entreprise Groupe Cegetel, Fondation Hasler, France T´el´ecom, Hitachi, ST Microelectronics, Swisscom, Texas Instruments, Thales

Abstract

It is well known that the peaks in log Mel-filter bank spectrum are important cues in characterizing the speech sounds. However, low energy perturbations in the power spectrum may become numerically significant after the log compression. We show that even if the spectral peaks are kept constant, the low energy perturbations in the power spectrum can create huge variations in the cepstral coefficients. We show, both analytically and experimentally, that exponentiating the log Mel-filter bank spectrum before the cepstrum computation can significantly reduce the sensitivity of the cepstra to spurious low energy perturbations. Mel-cepstrum modulation spectrum [3] is computed from the processed cepstra which results in further noise robustness of the composite feature vector. In experiments with speech signals, it is shown that the proposed technique based features yield a significant increase in speech recognition performance in non-stationary noise conditions when compared directly to the MFCC and RASTA-PLP features.

1 Introduction

As is well known, in the presence of commonly encountered additive noise levels, the formants are less affected as compared to the spectral "valleys" which exhibit spurious ripples. The DCT of a log Mel-filter bank spectrum (logMelFBS) which is commonly known as MFCC[2] feature vector, is sensitive to ripples in the spectral valleys which, otherwise, do not characterize the speech sounds. This is one of the reasons for the poor performance of MFCC features in additive noisy conditions. Observing that the higher amplitude portions (such as formants) of a spectrum are relatively less affected by noise, Paliwal proposed spectral subband centroids (SSC) as features [8, 9]. In this work, we analytically show that exponentiating the logMelFBS can decrease the sensitivity of the cepstra to the spurious perturbations in the logMelFBS valleys as compared to the peaks.

Lim has proposed the use of spectral root homomorphic deconvolution system (SRDS) [4] as an approximately more general case of logarithmic homomorphic deconvolution system (LHDS) [1]. SRDS uses a root compression $(.)^{\gamma},\ \gamma < 1$ of the mel-filter bank energies instead of the logarithmic compression used by LHDS. Although, Lockwood et. al [5] and Tokuda et. al [6] have proposed a unified approach to root Mel-cepstral coefficients (RMFCC), many researchers have used RMFCC with a motivation based on auditory and perceptual data. However, in this work, we use LHDS based MFCC features[2]. We provide a signal processing reason for the high sensitivity of the MFCC features towards additive noise and propose a solution to alleviate this problem by exponentiating the logMelFBS by a suitable positive power greater than unity. In [3], we proposed the use of Mel-cepstrum modulation spectrum (MCMS) features for robust ASR. MCMS features[3] are obtained by filtering cepstral trajectories using a bank of band-pass filters in the range $[2, 20]Hz$. In this work we derive MCMS features from the cepstra of the exponentiated logMelFBS. The experimental results show that these two sequential processing techniques synergistically improve the recognition rate in presence of additive non-stationary noise as compared to the MFCC and RASTA-PLP feature vectors.

2 Perturbations in log Mel-filter bank spectrum

One of the outcomes of logarithmic compression of the Mel-filter bank energies is the reduction of the dynamic range of the spectral amplitudes. Consequently, the spurious perturbations which are numerically insignificant in the power spectrum domain may become numerically significant after the logarithmic compression of the Mel-filter bank energies. In figure 1, we illustrate this problem. Blue and red curves are two instances of a logMelFBS with same formants but different perturbations in the low energy. These perturbations account for approximately $10^{-12}\%$ of the power spectral energy (before the log compression) and therefore do not characterize the speech sound. However, DCT being a linear transformation,

gives an equal weightage to the formants and the low energy filter bank outputs and therefore is sensitive to the spurious ripples. A natural solution to this problem is to weight the logMelFBS such that formants become more significant than the low energy mel-filter bank samples. To this end, a copy of the logMelFBS itself, is a good candidate for the "lifter" as it will emphasize the formants much more than the low energy log Mel-filter bank outputs. This is same as exponentiating the logMelFBS with a power $P,\ where\ P > 1$. In figure2, we plot squares of the two instances of the logMelFBS, same as in figure 1. As can be visually noted from the curves in figure 2, the formants have become more prominent as compared to the spurious ripple. In figure 3, the blue curve corresponds to the percentage absolute difference between the first 13 DCT coefficients of the two logMelFBS same as in figure 1 and red curve corresponds to the percentage absolute difference between the first 13 DCT coefficients of the squared logMelFBS same as in figure 2. The fact that the red curve lies below the blue curve, indicates that the squaring of the logMelFBS decreases the sensitivity of lower DCT coefficients towards spurious ripples in low energy region.

Consider k^{th} DCT coefficient of a N point sequence x. It can be approximately seen as a weighted sum of the "discrete" derivatives of the sequence X evaluated at k equidistant samples and multiplied by alternating signs. For instance, if $k = 5$ and $N = 10$, we have,

$$\begin{aligned} X_{DCT}(k) &= \textstyle\sum_{n=0}^{N-1} cos(\pi kn/N)x(n) \\ &= \textstyle\sum_{n=0}^{9} cos(\pi 5n/10)x(n) \\ = x(0)/2 + \textstyle\sum_{n=1}^{4}(-1)^n \frac{x(2n)-x(2n-2)}{(2n)-(2n-2)} &+ x(8)/2 \\ &\simeq \textstyle\sum_{n=1}^{4}(-1)^n x'(2n-1), \end{aligned} \tag{1}$$

where, $x'(n)$ denotes "discrete" derivative of x. Therefore the sensitivity of the DCT of the logMelFBS can be approximately measured in terms of the sensitivity of derivatives of the logMelFBS. We define the sensitivity index $\rho(a, b)$ as the ratio of derivatives of the function $log(x)$ at a Mel-formant energy $x = a$ and a low Mel-filter bank energy value $x = b$. Given (1), we expect $\rho(a, b)$ to measure the relative contributions of a peak of the logMelFBS and the low energy Mel-filter bank energies in a DCT coefficient which is a cepstral coefficient.

$$\begin{aligned} \rho(a,b) &= \tfrac{log'(x)|_{x=a}}{log'(x)|_{x=b}} = \tfrac{1/a}{1/b} \\ &= b/a\ where\ a \gg b \\ &\Rightarrow \rho(a,b) \ll 1.00 \end{aligned} \tag{2}$$

Similarly we define the sensitivity index $\sigma(a, b)$ as the ratio of the derivatives of the function $sign(log(x))[log(x)]^P$ at a Mel-formant energy $x = a$ and a low Mel-filter bank energy value $x = b$.

$$
\begin{aligned}
\sigma(a,b) &= \frac{P[sign(log(a))][log(a)]^{P-1}/a}{P[sign(log(b))][log(b)]^{P-1}/b} \\
&= \frac{[sign(log(a))][log(a)]^{P-1}}{[sign(log(b))][log(b)]^{P-1}}(b/a) \\
&= \frac{[sign(log(a))][log(a)]^{P-1}}{[sign(log(b))][log(b)]^{P-1}}\rho(a,b)\ where\ a \gg b \\
&\Rightarrow \sigma(a,b) > \rho(a,b)\ where\ a \gg b, P > 1.0
\end{aligned}
\tag{3}
$$

The value of $\rho \ll 1.0$ in (2) implies that a unit change in the low Mel-filter bank energy value, namely "b" will have a far greater influence on the computation of the DCT of logMelFBS as compared to a unit change in the Mel-formant energy, namely "a". Therefore, it can be seen in the light of (2) that the DCT of the logMelFBS is quite sensitive to the perturbations in the low-energy regions as compared to those around the formants. However, for the domain $1.0 \leq b \ll a < \infty$ and $P > 1$, $\sigma(a,b)$ is always greater than $\rho(a,b)$. This can be achieved by using $(log(x+1))^P$ as x being power spectral energy never takes negative values. The fact that the $\sigma(a,b)$ is always greater than the $\rho(a,b)$ implies that we have been able to decrease the sensitivity of cepstral coefficients to spurious low energy perturbations. An important parameter in the above mentioned processing scheme is the exponent P. As can be seen from (3), the sensitivity ratio $\sigma(a,b)$ increases exponentially as the exponent P increases. However, a large value of P will result in the case where the spectral modulations of the largest formant takes very high numerical values which render the spectral modulations of the other formants numerically insignificant relative to those of the largest formant. Therefore an intermediate value of P is the most suitable for such a processing scheme.[1]

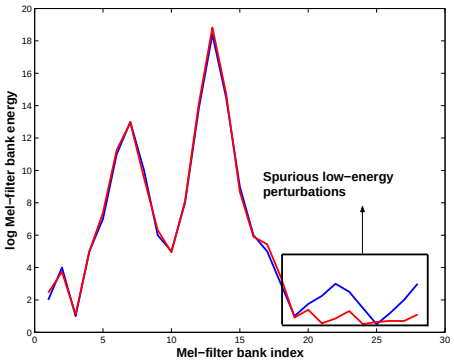

Figure 1: *Log Mel-filter bank energies of clean and noisy(perturbed) speech.*

[1]The experiments results with different values of P reconfi rmed these observations.

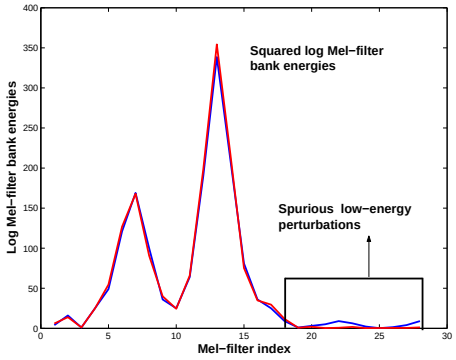

Figure 2: *Square of the log Mel-filter bank energies of clean and noisy(perturbed) speech.*

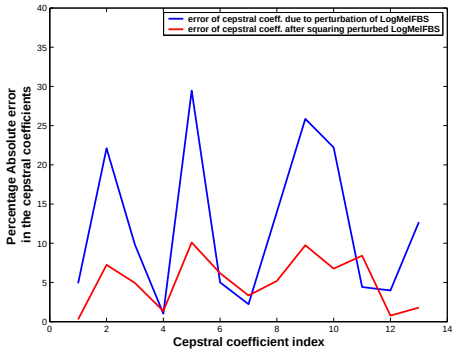

Figure 3: *Absolute percentage error between the cepstral coefficients due to perturbations. Blue curve corresponds to the DCT of the log Mel-filter bank spectrum while red curve corresponds to the DCT of the squared log Mel-filter bank spectrum.*

3 Experiments and Results

In order to assess the effectiveness of the proposed scheme for reducing the effect of spurious perturbations in the low Mel-filter bank energies, speech recognition experiments were conducted on the OGI Numbers95 corpus [11] using the proposed processing scheme for the logMelFBS. The lexicon size for this connected digits recognition task is 30 words with 27 different phonemes. To verify the robustness of the features to noise, the clean test utterances were corrupted using additive non-stationary *factory* noise and *f16 cockpit* noise from the Noisex92 [12] database. Throughout the experiments, Mel-frequency cepstral coefficients (MFCC) [2] and their temporal derivatives have been used as speech features. Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) based speech recognition systems were trained using public domain software HTK [9] on the clean training set from the original Numbers95 corpus. The system consisted of

80 tied-state triphone HMM's with 3 emitting states per triphone and 12 mixtures per state. Three kinds of feature sets were generated:

- [MFCC+Deltas:] 13 MFCCs with deltas.
- [RMFCC+Deltas: generated by root Mel-filter bank spectrum with $R = 0.10$] 13 root Mel-cepstral coefficients with deltas.
- [ExpoMFCC+Deltas: generated by exponentiated logMelFBS with $P = 2.7$] 13 exponentialted log-Mel-cepstral coefficients with deltas.

Per utterance cepstral mean subtraction was applied to each of the above feature vectors. The speech recognition results using the above mentioned feature sets in clean and noisy conditions are reported in table 1. The root $R = 0.10$ and the exponent $P = 2.7$ gave the best recognition results for the RMFCC and ExpoMFCC features respectively. The exponentiated logMelFBS MFCC system performs significantly better than the usual MFCC features in the noisy conditions. We note that the performance of the proposed features is similar to that of RMFCC features using the optimal value of the root $R = 0.10$. Figure 4 illustrates the fact that the proposed technique can significantly reduce the mismatch between clean and noisy MFCC features.

In [3], we proposed the use of Mel-cepstrum modulation spectrum (MCMS) features for robust ASR. MCMS features[3] are obtained by filtering cepstral trajectories using a bank of band-pass filters in the range $[2, 20]Hz$. In this work we derived MCMS features from the cepstra of the exponentiated logMelFBS. The recognition results are reported in table 2. All the features in this table have mean and variance normalized cepstra. The superior performance of ExpoMFCC+MCMS features can be noticed in the last column of the table 2. The average word error rate (WER) for the ExpoMFCC+MCMS features in clean and all the noisy conditions in 15.8%. This corresponds to a relative improvement of 24.0% over RASTA-PLP features and 11.4% over the optimal RMFCC features.

Table 1: *Word error rate results for factory and f16 noise. The best results for RMFCC (R=0.10) and Exponentiated MFCC (P=2.7) are reported.*

SNR	MFCC	RMFCC	**ExpoMFCC**
Clean	6.1	6.1	6.2
Fact SNR 12	14.0	12.0	11.6
Fact SNR 6	31.5	20.6	20.3
Fact SNR 0	75.7	45.7	44.3
F16 SNR 12	15.8	12.3	12.1
F16 SNR 6	32.8	20.8	20.9
F16 SNR 0	75.1	44.2	43.4

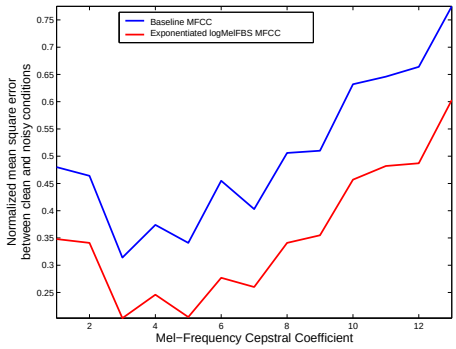

Figure 4: *Mean square error of MFCC vectors in clean and noisy conditions, normalized by the average power of the corresponding MFCC feature vector in clean condition. Blue curve corresponds to baseline MFCC while red curve corresponds to MFCC derived by squaring the log Mel-filter bank spectrum. These mean estimates were computed using nearly 160000 speech frames.*

Table 2: *Word error rate results for factory and f16 noise. All the features in this case have cepstral mean and variance normalization.*

SNR	RASTA-PLP	RMFCC	**ExpoMFCC+MCMS**
Clean	6.5	6.1	5.0
Fact SNR 12	10.6	10.4	9.2
Fact SNR 6	18.4	16.7	15.2
Fact SNR 0	37.9	35.3	31.6
F16 SNR 12	11.2	10.2	9.5
F16 SNR 6	17.9	15.7	14.4
F16 SNR 0	34.8	28.9	26.0
Average	19.6	17.6	15.8

4 Conclusion

We identify a numerical sensitivity problem with the MFCC[2] features. It is analytically shown that by exponentiating the logMelFBS one can desensitize the MFCC coefficients to spurious low-energy spectral perturbations. Finally, Mel-cepstrum modulation spectrum[3] is derived from the cepstra which in turn has been derived by exponentiating the logMelFBS. The experimental results show that significant noise robustness can be achieved by the use of the proposed features in all conditions as compared to the RASTA-PLP and root MFCC features.

References

[1] A. V. Oppenheim and R. W. Schafer, Discrete-Time Signal Processing, pp. 771-772, Prentice-Hall, N.J., USA, 1989.

[2] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences, " IEEE Trans. on ASSP, Vol. ASSP-28, No. 4, August 1980.

[3] V. Tyagi, I. McCowan, H. Bourlard and H. Misra, " Mel-Cepstrum Modulation Spectrum Features (MCMS) for Robust ASR, ", In the Proc. of IEEE ASRU 2003, US Virgin Islands, USA, 2003.

[4] J. S. Lim, " Spectral Root Homomorphic Deconvolution system, " IEEE Trans. on ASSP, Vol. ASSP-27, No. 3, June 1979.

[5] P. Alexandre and P. Lockwood, " Root Cepstral Analysis: A unified view. Application to speech processing in car noise environments, " Speech Communication, Vol.12, pp:277-288, 1993.

[6] K. Tokuda, T. Kobayashi, T. Masuko and S. Imai, " Mel-Generalized Cepstral Analysis - A Unified Approach to Speech Spectral Estimation, " Proc. of the IEEE-ICASSP, 1994.

[7] J. R. Deller, J. G. Proakis and J. H. L. Hansen, Discrete Time Processing of Speech Signals, pp. 377-378, Macmillan Publishing Company, New York, USA, 1993

[8] J. Chen, Y. Huang, Q. Li, and K. K. Paliwal, " Recognition of Noisy Speech Using Dynamic Spectral Subband Centroids, " IEEE Signal processing Letters, Vol. 11, No. 2, February 2004.

[9] K. K. Paliwal, "Spectral Subband centriod features for speech recognition, " in Proc. ICASSP, Vol. 2, 1998, pp. 617-620.

[10] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK Book, Cambridge University, 1995.

[11] R. Cole, M. Noel, T. Lander, and T. Durham, " New telephone speech corpora at CSLU," in Proc. of European Conference on Speech Communication and Technology, 1995, vol.1, pp.821-824.

[12] A. Varga, H. Steeneken, M. Tomlinson and D. Jones, " The NOISEX-92 study on the effect of additive noise on automatic speech recognition, " Technical report, DRA Speech Research Unit, Malvern, England, 1992.