ON DE-EMPHASIZING THE SPURIOUS COMPONENTS IN THE SPECTRAL MODULATION FOR ROBUST SPEECH RECOGNITION

Vivek Tyagi and Christian Wellekens

Institute Eurecom
B.P 193 -06904 Sophia Antipolis, France
Vivek.Tyagi@eurecom.fr and Christian.Wellekens@eurecom.fr

ABSTRACT

It is well known that the peaks in log Mel-filter bank spectrum essentially represent the "formants" of the speech signal and are important cues in characterizing the sound. However, the perturbations in the low energy log Mel-filter bank spectrum create unnecessary sensitivity in the cepstral comparison, especially in the presence of the additive noise. In this paper, we present a technique to suppress this unnecessary sensitivity of the log Mel-filter bank spectrum (logMelFBS) of the speech signals, while preserving the fundamental formant structure. From the practical point of view, our technique is quite similar to the spectral root homomorphic deconvolution systems (SRDS) [3]. However, we work with log homomorphic deconvolution system (LHDS) [1] and use an exponentiation of logMelFBS to emphasize the spectral peaks (formants). 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 features, while achieving slightly better performance in clean conditions. The proposed technique yields almost similar performance as compared to the root Mel-cepstral coefficients (RMFCC) in the noisy as well as clean conditions.

1. INTRODUCTION

Automatic speech recognition (ASR) typically involves windowing speech signal into 20ms to 30 ms long segments. The resulting short time power spectrum estimate is filtered by a bank of mel-filters and then compressed by a logarithmic non-linearity. A DCT operation on the log Mel-filter bank spectrum (logMelFBS) and retaining the lower DCT coefficients yields a smoothed spectral modulation estimate which is well known as Mel frequency cepstral coeffients (MFCC) [2].

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 logMelFBS, which is a MFCC feature vector, essentially estimates spectral modulations and 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 [7, 8]. In this work, we will investigate the use of logMelFBS exponentiation to increase the sensitivity of the logMelFBS DCT coefficients towards the formants as compared to the spurious perturbations in the logMelFBS valleys.

Lim has proposed the use of spectral root homomorphic deconvolution system (SRDS) [3] 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. Many researchers have used SRDS to obtain root Mel-cepstral coefficients (RMFCC), which have been shown to be superior to MFCCs in clean and noisy conditions [4], [5]. However, in this work, we use LHDS based MFCC features[2]. We investigate a plausible reason for the high sensitivity of the logMelFBS towards additive noise and propose a solution to alleviate this problem by exponentiating the logMelFBS by a suitable positive power. The experimental results show the efficacy of the proposed technique as compared to the MFCC feature vectors.

2. PERTURBATIONS IN LOG MEL-FILTER BANK SPECTRUM

Theoretically speaking, the logarithm of the Mel-filter bank spectrum is used for homomorphic deconvolution of speech signal into the power spectral envelope and the excitation spectra [1]. However, in practice, 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, may become numerically significant after the logarithmic compression of the mel-Filter bank energies. In figure 1, we illustrate the problem of spurious perturbations in logMelFBS which become numerically significant in the computation of the lower DCT coefficients. Blue curve corresponds to a "clean" logMelFBS with two formants, while the red curve corresponds to a noisy (perturbed) logMelFBS. We note from the red curve that, the formants of the perturbed logMelFBS are relatively unchanged, while there is a spurious ripple in the low energy region. 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 square of the clean logMelFBS and square of the perturbed logMelFBS which are the 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 9 DCT coefficients of the original and the perturbed logMelFBS as in figure 1 and red curve corresponds to the percentage absolute difference between the first 9 DCT coefficents of the squared original and the squared perturbed logMelFBS 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.

In figures 4, 5 and 6, we plot graphs of $log(x)$ and $sign[log(x)][log(x)]^2$ in the domains $x \in [10,\ 1000]$, $x \in [0.1,\ 1]\, and\, x \in [0.001,\ 0.1]$ which corresponds to the compression, the compression-cum-expansion and the expansion regions of the logarithmic function, respectively. Consider k^{th} DCT coefficient of a N point sequence x, which can be approximately seen as a weighted sum of the "discrete" derivatives of the sequence X evaluated at k equidistant samples. For instance, if $k = 5$ and $N = 10$, we have,

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

where, $x'(n)$ denotes "discrete" derivative of x. Therefore the sensitivity of 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 $x = b$.

$$\begin{aligned} \rho(a, b) &= \frac{1/a}{1/b} \\ &= b/a \; where \; a \gg b \\ &\Rightarrow \rho(a, b) \ll 1.00 \end{aligned} \tag{2}$$

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 $x = b$.

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

The value of $\rho \ll 1.0$ 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 (1) that the DCT of logMelFBS is quite sensitive to the perturbations in the low-energy regions as compared to those around the formants. For the domain $1.0 \leq b \ll a < \infty$ and $P > 1$, $\sigma(a, b)$ is always greater than $\rho(a, b)$. However for the values of $b \in [0, 1.0]$ and $b \ll a$, $\sigma(a, b)$ can take values less than $\rho(a, b)$. For instance, if $a = 10^2$ and $b = 10^{-5}$, then $\sigma(a, b) = (2/5)^{P-1}\rho(a, b)$. Therefore, if we threshold the energies of the Mel-filter bank which are below 1.0 to a constant value equal to 1.0 and then take the P^{th} power of the logarithm of the Mel-filter bank energies, followed by a DCT, we can increase the sensitivity ratio from $\rho(a, b)$ to $\sigma(a, b)$. This will emphasize the formant spectral modulations while de-emphasizing the spurious contribution of the spectral modulations in the very low-energy Mel-spectrum. Therefore, the entire $x - axis$ in the figure 4 and all the points to the right of $point\ A$ in figure 5 constitute the desirable domain of operation. The entire $x - axis$ to the left of $point\ A$ is mapped to $point\ A$. In order to avoid thresholding of a significant proportion of the speech signals, all the train and test utterances can be multiplied by a common scale factor such that the average power of the utterances is above 20db. This is not such a severe restriction. The need for scaling will arise if the speech signal has been artificially scaled down such that it is even inaudible to humans. For instance, in our experiments we did not have to use any scaling. 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 will assume such high numerical values that the spectral modulations of the other formants will become numerically insignificant relative to those of the largest formant. Therefore an intermediate value of P is the most suitable for such a processing scheme. The experimental results reported in this paper have reconfirmed these observations.

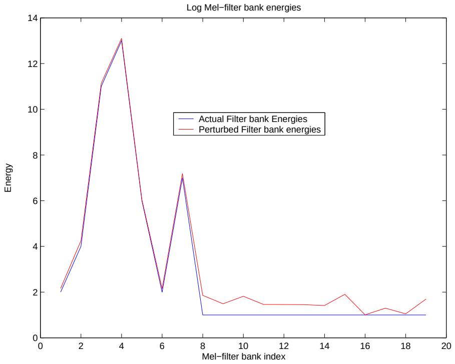

Fig. 1. *Log Mel-filter bank energies of clean and noisy(perturbed) speech.*

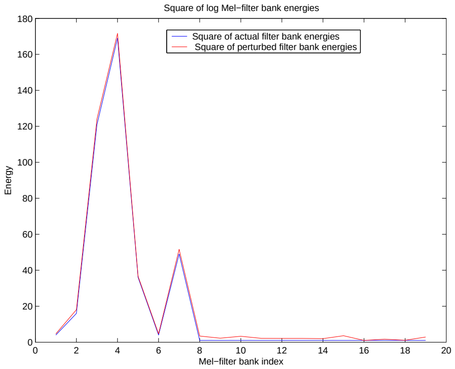

Fig. 2. *Square of the log Mel-filter bank energies of clean and noisy(perturbed) speech.*

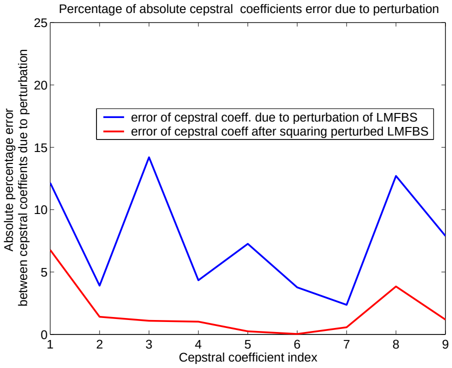

Fig. 3. *Absolute percentage error between the cepstral coefficients due to perturbations. Blue curve corresponds to the DCT of the log Mel-filter bank spectrum while red curve corresponds to the DCT of the squared log Mel-filter bank spectrum.*

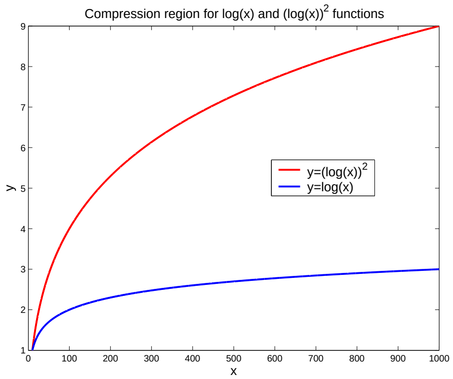

Fig. 4. *Compression region of* $[log(x)]$ *and* $[log(x)]^2$ *functions*

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 [10] 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 from the Noisex92 [11] 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 [8] 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= 0.06, 0.08 and 0.12] 13 root Mel-cepstral coefficients with deltas.
- [ExpoMFCC+Deltas: generated by exponentiated logMelFBS with P=1.5, 2.0 and 3.0] 13 exponentialted log-Mel-cepstral coefficients with deltas.

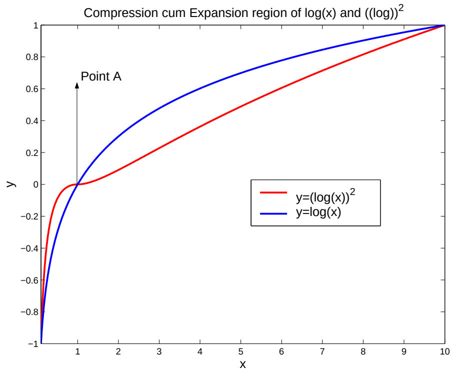

Fig. 5. *Compression cum Expansion region of* $[log(x)]$ *and* $[log(x)]^2$ *functions*

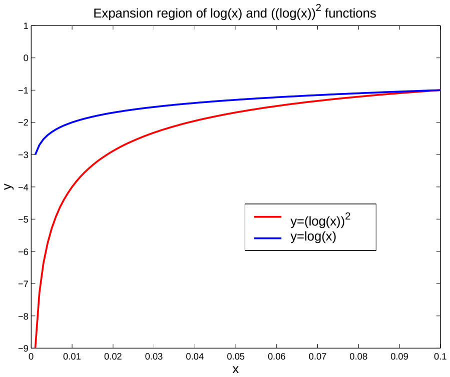

Fig. 6. *Expansion region of* $[log(x)]$ *and* $[log(x)]^2$ *functions*

The power spectrum in each of the above feature vectors was calculated using a Hamming window of length 37.5ms with a shift of 10ms. Finally, per utterance cepstral mean substraction 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 exponentiated log-MelFBS MFCC system with $P = 2.0$ performs significantly better than the usual MFCC features in the noisy conditions and also brings about certain improvement in the clean conditions. We note that the performance improvement over the baseline MFCC feature vector, starts to drop as the power P is increased beyond the value 2.0. This is consistent with the discussion in section 2 where we argue that as the power P increases, the spectral modulations of the largest formant will assume such high numerical values that the spectral modulations of the other formants will become numerically insignificant relative to those of the largest formant. We note that the performance of the proposed features is similar to that of RMFCC features using the best value of the root $R = 0.08$. This value is similar to the one reported in [5].

In figure 7, the blue curve represents the mean square error between the MFCC feature vector computed from a clean speech frame and the same frame corrupted by an additive non-stationary factory noise at SNR12, followed by a normalization by the average power of MFCC feature vector in clean condition. Whereas, the red curve represents the mean square error for MFCC feature vector, computed by exponentiating the logMelFBS with power $P = 2$. As above, the mean square error of the processed MFCC feature vectors in the clean and the noisy conditions, is normalized by the average power of the processed MFCC feature vector in the clean conditions. These average estimates were computed using 16,000 speech frames in clean condition and their noisy instances were obtained by adding the non-stationary factory noise at SNR12. Therefore on an average, the proposed technique significantly reduces the mismatch between clean exponentiated MFCC vectors and their additive noise corrupted versions as compared to baseline MFCC feature vector.

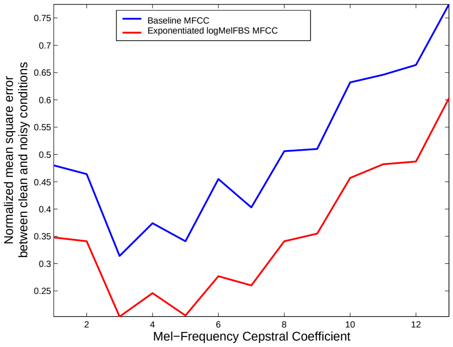

Fig. 7. *Mean square error of MFCC vectors in clean and noisy conditions, normalized by the average power of the corresponding MFCC feature vector in clean condition. Blue curve corresponds to baseline MFCC while red curve corresponds to MFCC derived by squaring the log Mel-filter bank spectrum.*

Table 1. *Word error rate results for factory noise. The best results for RMFCC (R=0.08) and Exponentiated MFCC (P=2.0) are in italics*

feature	Clean	SNR12
MFCC	6.3	14.4
RMFCC Root=0.06	6.0	12.7
RMFCC Root=0.08	6.0	12.1
RMFCC Root=0.12	6.5	12.4
Exponentiated logMelFBS MFCC P=1.5	6.3	11.9
Exponentiated logMelFBS MFCC P=2.0	6.0	11.7
Exponentiated logMelFBS MFCC P=3.0	6.7	12.5

4. CONCLUSION

We have shown that emphazing the peaks in the logMelFBS using a “filter” which is a copy of the logMelFBS can significantly suppress the spurious perturbations in the low-energy logMelFBS. Consequently, the lower DCT coefficients of the exponentiated logMelFBS become less sensitive to the spurious low-energy perturbations, which otherwise can significantly impair the performance of a MFCC feature based HMM-GMM speech recognizer. We identitify, the different domains of the $log(x)$ and $sign[log(x)][log(x)]^P$ functions and their relation to sensitivity index ρ and σ and the exponentiation power P. Finally, the proposed technique is compared to the root-cepstral coefficient (RMFCC) features [3],[4],[5]. Sarilaya’s and Hansen’s motivation for the use of RMFCC is to reduce the real-time processing factor (RTF)[5]. However, in this work, our motivation is to alleviate the numerical sensitivity problem of the logarithmic homomorphic deconvolution systems (LHDS), by exponentiating the logMelFBS. Experiments indicate that both of the techniques yield similar results.

5. 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] J. S. Lim, “ Spectral Root Homomorphic Deconvolution system, ” IEEE Trans. on ASSP, Vol. ASSP-27, No. 3, June 1979.

[4] 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.

[5] R. Sarikaya and J. H. L. Hansen, “ Analysis of root-cepstrum for acoustic modeling and fast decoding in speech recognition, ”, In the Proc. of Eurospeech 2001, Aalborg, Denmark, Sept. 2001.

[6] 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

[7] 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.

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

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

[10] 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.

[11] 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.