AN ALGORITHM FOR ROBUST SIGNAL MODELLING IN SPEECH RECOGNITION

Rivarol Vergin

CML Technologies Inc.
75 Blvd de la technologie, Hull, Quebec, Canada
email: rvergin@cmltech.com

ABSTRACT

The most popular set of parameters used in recognition systems is the mel frequency cepstral coefficients. While giving generally good results, it remains that the filtering process, as used in the evaluation of these parameters, reduces the signal resolution in the frequency domain, which can have some impact in discriminating between phonemes. This paper presents a new parameterization approach that preserves most of the characteristics of mel frequency cepstral cofficients while maintaining the initial frequency resolution obtained from the fast Fourier transform. It is shown, by the results obtained, that this technique can significantly increase the performance of a recognition system.

1. INTRODUCTION

The first step of a continuous speech recognition process is parameterization, whose role is data reduction in converting the input signal into parameters while preserving virtually all of the speech signal information dealing with the text message. Common parameters are LPC coefficients, energies in a channel vocoder, and mel frequency cepstral coefficients [1] (MFCC), which remain the most popular set of parameters used in speech recognition systems.

The evaluation technique of the MFCCs involves many steps: fast Fourier transform, filtering, and cosine transform. Following the dynamic range effects of the ear, the filtering procedure results in a representation of the spectral energy on the mel scale through the use of a set of filters, generally 24, equally spaced at low frequency and continuously increasing beyond 1 kHz.

The idea consisting of mapping an acoustic frequency to a perceptual frequency scale is the most important aspect of the mel frequency cepstral coefficients. But obviously the ill effect of this technique is a reduction of the frequency resolution inherent in the filtering process. This reduction of the frequency resolution can have some impact in discriminating between phonemes and consequently on results obtained with a continuous speech recognition system.

This paper shows how the filtering process combined with the cosine transform, occuring in the evaluation of the mel frequency cepstral coefficients, leads to an un-harmonic development. Based on this new representation of the MFCC, we suggest in this paper a new method of evaluation of the input parameters that respects the fundamental concept of the mel scale while keeping unchanged the initial frequency resolution obtained from the fast Fourier transform. Results obtained with these new coefficients give a confidence interval for their use in continuous speech recognition systems.

2. MEL FREQUENCY CEPSTRAL COEFFICIENTS

Most useful parameters in speech processing are found in the frequency domain, because the vocal tract produces signals that are more consistently and easily analyzed spectrally than in the time domain. Repeated utterances by one speaker of a sentence often differ considerably in the time domain while remaining quite similar in the frequency domain. For these reasons, spectral analysis is used primarily to extract relevant parameters from speech signals. One of the most popular sets of parameters used in recognition systems, the mel frequency cepstral coefficients [1], is evaluated according to the same principal.

2.1. Evaluation technique of MFCC

First, a fast Fourier transform is calculated using a 30 ms Hanning window. Assuming that $X = [x_1, \cdots, x_K]$ is a spectral energy vector, the second step is the evalaution of the channel energy vector, $E = [e_1, \cdots, e_J]$, where each element, e_j, is given by:

$$e_j = \sum_{k=1}^{K} \phi_j(k) x_k; \quad 1 \leq j \leq J \tag{1}$$

Generally, $J = 24$ and $K = 256$. ϕ_j is a set of triangular filters whose centers, see Picone [2], are equal to: [100, 200,

300, 400, 500, 600, 700, 800, 900, 1000, 1149, 1320, 1516, 1741, 2000, 2297, 2639, 3031, 3482, 4000, 4595, 5278, 6063, 6964, 8000] Hz. This particular spacing between successive filters center, allows one to map the acoustic frequency to a perceptual frequency scale, that is, the mel scale, hence the name of these coefficients. But clearly, according to equation (1), the filtering process reduces the initial frequency resolution obtained from the fast Fourier transform. This reduction of the frequency resolution can have some impact in discriminating between phonemes and consequently on the performance of a recognition system.

To see how this problem can be overcome while maintaining the mapping of the acoustic frequency to the perceptual frequency scale, it is necessary to consider the last step involved in the evalaution of the MFCC. It consists of a cosine transform of the log channel energy vector, resulting in a set of coefficients,

$$c_m = \beta \sum_{j=1}^{J} \cos(m\frac{\pi}{J}(j-0.5)) \log_{10}(e_j). \qquad (2)$$

The amplification factor, β, accomodates the dynamic range of the coefficients c_m.

This last equation can also be viewed as a scalar product between the vector of channel energy, E, and a set of vectors, W_m, whose elements,

$$w_{m,j} = \cos(m\frac{\pi}{J}(j-0.5)) \quad 1 \leq j \leq J, \qquad (3)$$

are located at the same position in the frequency domain as the energy elements, e_j. As an example, the figure 1 shows the distribution of the elements of W_5 in the frequency domain. It can be observed that the equal spacing of harmonic series is not respected, this is the positive effect of the mel scale. The negative effect is the relatively large distance between successive points that form the curve highlighting the reduction of frequency resolution above-mentioned. The next section suggests an algorithm to solve this problem.

3. HIGH FREQUENCY RESOLUTION COEFFICIENTS

Obviously, the crucial element in figure 1 is the global form of the curve. From this representation, we can conclude that the combination of the mel scale with the cosine series leads to a specific set of curves acting as projection base for the channel energy vector E.

We suggest in this section an algorithm which leads to a new set of vectors $\widetilde{W}_m$ acting as a projection base for the vector of energy X, obtained after the fast Fourier transform, instead of the channel energy vector E. This algorithm allows us to avoid the problem of reduction of frequency resolution associated with E. The global form of the curves obtained when the vectors $\widetilde{W}_m$ are represented in the frequency domain is basically same as those obtained with the vectors W_m. The major difference comes from the fact that the vectors $\widetilde{W}_m$ contain K elements, with $K = 256$; this is contrary to the vectors W_m that contain J elements, with $J = 24$.

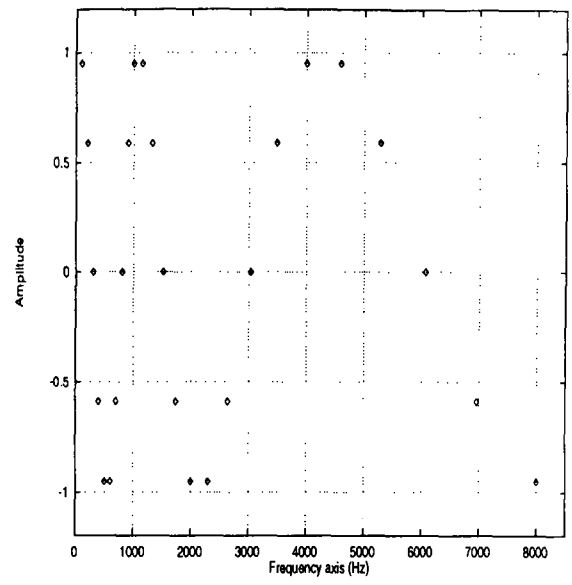

Figure 1. Representation in the frequency domain of W_5.

3.1. Evaluation technique

The suggested algorithm involves the following steps, for each value of m:

1. Evaluate a vector of position P_m, whose elements, $P_{m,l}$, are given by:

$$P_{m,l} = 700(10^{\frac{l}{m}\frac{\theta}{2595}} - 1); \quad 0 \leq l \leq m. \qquad (4)$$

The right side of this equation is based on the inverse of the equation, see [2] [3], defining the mel scale. The value $\theta = 2840$, is chosen in such a way that the last element, $P_{m,m}$, is equal to 8000 Hz.

2. Define between each pair of elements, $P_{m,l}$ and $P_{m,l+1}$, a subset of vectors, $\widetilde{w}_{m,l}$, given by:

$$\widetilde{w}_{m,l} = \{(-1)^l \cos(\frac{\pi}{I_l}(i-0.5)) \quad | \quad 1 \leq i \leq I_l\}, \quad (5)$$

where I_l is the total number of energy elements x_k between $P_{m,l}$ and $P_{m,l+1}$.

3. Concatenate all the vectors $\widetilde{w}_{m,l}$ for $0 \leq l \leq m$ to obtain the final vector $\widetilde{W}_m$, that is,

$$\widetilde{W}_m = \bigcup_{l \in [0,m-1]} \widetilde{w}_{m,l}. \quad (6)$$

The final vector $\widetilde{W}_m$ contains the same number of elements as the initial spectral energy vector X. The curve, shown in figure 2, represents $\widetilde{W}_5$ in the frequency domain. It can be observed that the global form of the curve in figure 2 is similar to the one in figure 1, while beeing constituted by a greater number of points.

The new set of coefficients, c^*_m, obtained using $\widetilde{W}_m$ is given by:

$$c^*_m = \beta \sum_{k=1}^{K} \widetilde{W}_{m,k} \log_{10}(x_k). \quad (7)$$

Because vectors $\widetilde{W}_m$ contain K elements, vector X is used instead of vector E in the evaluation of c^*_m. It is then possible to conclude that the frequency resolution involved in the evaluation of c^*_m is similar to that obtained after the fast Fourier transform. In our experiments the first spectral energy element, x_1, has been not used, which cancels the effect of the DC component of the input signal; a Gram-Schmidt orthogonalization procedure has been used to obtain an orthogonal base from the set of vectors $\widetilde{W}_m$; and the spectral vector X have been compensated, using the technique described in [4], to reduce the effect of high concentration of energy at low frequency.

4. COMPARATIVE RESULTS

Experiments described in this section have been conducted using the INRS speech recognizer [5], which is a large-vocabulary speaker-independent continuous speech recognition system. The system accepts at the input a set of 15 coefficients and their first-order derivatives evaluated every 10 ms using a Hanning window of 30 ms. The output of the recognizer is a succession of words that match the input speech data according to the language model, a bigram model.

The speech corpus used in these experiments came from ATIS corpora, with a vocabulary of 1087 words. 285 speakers have been used for the training and 10 for the tests. Males and females are present in both training and testing sets. Table 1 summarizes the results obtained. The coefficients c are classical mel-frequency cepstral coefficients and coefficients c^* are the new set of coefficients suggested in this paper. It can be observed that results are better with c^* than with c for male as well as for female speakers. This lets us believe that the suggested coefficients c^* can increase the performance of a recognition system.

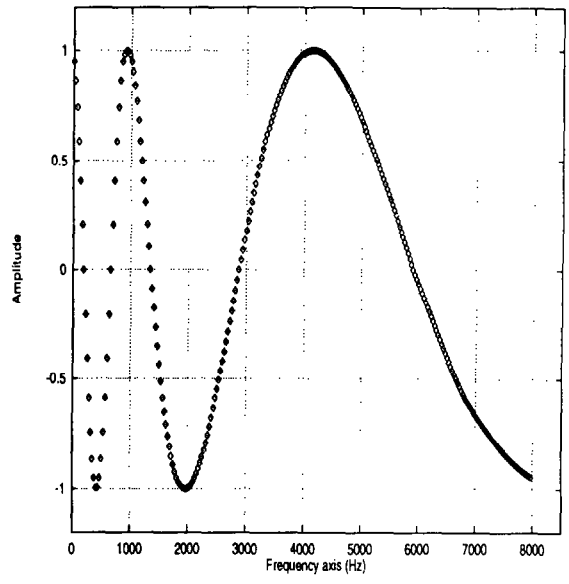

Figure 2. Representation in the frequency domain of $\widetilde{W}_5$.

type of coefficients	word accuracy (%)		
	male	female	average
c	88.38	85.24	86.69
c^*	91.54	86.05	88.42

Table 1. Comparative results for coefficients c and c^*.

5. SUMMARY

In this paper we have examined one of the most used sets of parameters, the mel frequency cepstral coefficients. We have shown that the filtering process, involved in the evaluation of these coefficients, reduces the initial frequnecy resolution obtained from the fast Fourier transform. It has been also shown that the filtering process combined with the cosine transform leads to an un-harmonic development in the frequency domain. This new representation of MFCC has been used to evaluate a new set of parameters that respects the human perceptual scale while keeping unchanged the initial frequency resolution.

Comparative results between mel frequency cepstral coefficients and the new set of coefficients proposed, confirm our assumption that a reduction in the frequency resolution can have some negative effects on the performance of a continuous speech recognition system.

REFERENCES

[1] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition" IEEE Trans. Acoust. Speech and Signal Processing, vol. ASSP-28, pp. 357-366, 1980.

[2] J. W. Picone, "Signal Modeling Techniques in Speech Recognition" Proc. IEEE, vol. 81, No. 9, pp 1215-1247, 1993.

[3] D. O'Shaughnessy, "Speech Communications: Human and Machine", Reading, MA: Addison-Wesley, 1987.

[4] R. Vergin, D. O'Shaughnessy, V. Gupta, "Compensated Mel frequency Cepstrum Coefficients", Proc. IEEE Int. Conf. ASSP, may 1996.

[5] P. Kenny, R. Hollan, G. Boulianne, H. Garudadri, Y.M. Cheng, M. Lennig, D. O'Shaughnessy, "Experiments in Continuous Speech Recognition with a 60,000 Word Vocabulary", Int. Conf. on Speken Language Processing, Banff, Can.,pp. 225-228, 1992.