AUTOMATIC SPEECH RECOGNITION BASED ON CEPSTRAL COEFFICIENTS AND A MEL-BASED DISCRETE ENERGY OPERATOR

Hesham Tolba *Douglas O'Shaughnessy*

INRS-Télécommunications, Université du Québec
16 Place du Commerce, Verdun (Île-des-Soeurs),
Québec, H3E 1H6, Canada
{tolba, dougo}@inrs-telecom.uquebec.ca

ABSTRACT

In this paper, a novel feature vector based on both Mel Frequency Cepstral Coefficients (MFCCs) and a Mel-based nonlinear Discrete-time Energy Operator (MDEO) is proposed to be used as the input of an HMM-based Automatic Continuous Speech Recognition (ACSR) system. Our goal is to improve the performance of such a recognizer using the new feature vector. Experiments show that the use of the new feature vector increases the recognition rate of the ACSR system. The HTK Hidden Markov Model Toolkit was used throughout. Experiments were done on both the TIMIT and NTIMIT databases. For the TIMIT database, when the MDEO was included in the feature vector to test a multi-speaker ACSR system, we found that the error rate decreased by about 9.51%. On the other hand, for NTIMIT, the MDEO deteriorates the performance of the recognizer. That is, the new feature vector is useful for clean speech but not for telephone speech.

1. INTRODUCTION

In this paper, we introduce a novel combination of features to be used as the output of the front-end analyzer of an **ACSR** system. The new element that we combine with the **MFCC** coefficients is the Teager Energy. Teager Energy calculation is based on the fact that the speech signal can be modeled as the sum of N AM-FM signals. This model represents each component of the speech signal as a signal with a combined amplitude modulation (AM) and frequency modulation (FM) structure. Hence, we can apply Teager's algorithm [1] for computing the energy of a signal. We use this algorithm as the basis for a new energy measure that replaces the traditional energy measure; it is used for a new time-frequency feature vector for speech recognition.

The nonlinear energy operator first developed by Kaiser [1] and its discrete-time counterpart have found several applications in the speech processing area [2],[3]. This discrete-time energy operator is defined as $\Psi[x(n)] \triangleq x^2(n) - x(n-1)x(n+1), n = 0, 1, .., N-1$. In [2] it has been shown that, when the energy operator Ψ is applied to an AM-FM signal, it can approximately estimate the squared product of the amplitude and frequency signals.

Applying the energy operator to a speech signal to get the new Energy's parameter of the feature vector came from the fact that the amplitude of the speech signal sample is always dependent on its frequency and that the traditional energy measure reflects only the amplitude of the signal, whereas the energy operator reflects the variations in both amplitude and frequency of the speech signal. This fact motivated us to include this element in the input feature vector to an automatic speech recognition system to enhance its performance.

This paper will be organized into the following sections. The second section will present an introduction about the AM-FM Modulation Model, the **DEO**, spectral analysis, the cepstral coefficients and the **MFCCs**. Following this, the third section will discuss how the **MFCCs** and the **MDEO** could be combined to be used as the input feature vector of an **ACSR**. Experimental results that demonstrate the effectiveness of adding the **MDEO** in the feature vector are presented in section 4. Finally, in section 5 we conclude and discuss our present and future work.

2. BACKGROUND

2.1. AM-FM Modulation Model

Motivated by several nonlinear and time-varying phenomena during speech production, Maragos, Quatieri and Kaiser [2] proposed an AM-FM modulation model that represents each single speech resonance (formant) as an AM-FM signal. This model represents each resonance of a speech signal as a signal with a combined amplitude modulation (AM) and frequency modulation (FM) structure. Then, the speech signal $x(t)$ is modeled as the sum of N such AM-FM signals, one for each formant, as follows:

$$x(t) = \sum_{i=1}^{N} a_i(t) cos\big(2\pi[f_{c,i}t + \int_0^t q_i(\tau)\, d\tau] + \theta_i\big), \qquad (1)$$

where $f_{c,i}$ is the center value of the i^{th} formant frequency, $q_i(t)$ is the frequency modulating signal, and $a_i(t)$ is the time-varying amplitude. The i^{th} instantaneous formant frequency signal is $f_{inst,i}(t) = f_{c,i} + q_i(t)$. In the discrete-time domain the i^{th} discrete-time AM-FM signal is defined as

$$x_i(n) = a_i(n)\, cos\Big(\Omega_{c,i}\, n + \Omega_m \int_0^n q_i(k)\, dk\Big), \qquad (2)$$

where $a_i(n)$ is the discrete-time amplitude envelope, $\Omega_{c,i}$ is the carrier frequency and Ω_m is the modulation frequency. The digital instantaneous frequency of the discrete-time AM-FM signal, $\Omega_{inst,i}$, is defined as

$$\Omega_{inst,i}(n) = \Omega_c + \Omega_m q(n). \quad (3)$$

2.2. Discrete-Time Energy Operator

The nonlinear **DEO**, $\Psi[x(n)]$, first developed by Kaiser [1] and its discrete-time counterpart have found several applications in the speech processing area [2], [3]. The **DEO** $\Psi[x]$, which tracks the energy of a source producing an oscillation signal $x(t)$, is defined as

$$\Psi[x(t)] = [\dot{x}(t)]^2 - x(t)\ddot{x}(t), \quad (4)$$

where $\dot{x} = dx/dt$. In the discrete-time domain $\Psi[x(n)]$ is defined as:

$$\Psi[x(n)] \triangleq x^2(n) - x(n-1)x(n+1), n = 0, .., N-1. \quad (5)$$

In [2] it was shown that, when the energy operator Ψ is applied to an AM-FM signal, it can approximately estimate the squared product of the amplitude and frequency signals; i.e.,

$$\Psi[x(t)] \approx [a(t)\omega_{inst}(t)]^2, \quad (6)$$

assuming that $a(t)$ and $\omega_{inst}(t)$ do not vary too fast with time compared to the carrier frequency ω_c. In the discrete-time domain $\Psi[x(n)]$ could be written as:

$$\Psi[x(n)] \approx [a(n)\Omega_{inst}(n)]^2. \quad (7)$$

From Equation (7) we can see that the energy operator is a function of both amplitude and frequency of the signal samples. Applying the energy operator to a speech signal to classify it came from the fact that the amplitude of the speech signal sample is always dependent on its frequency and that the traditional energy measure reflects only the amplitude of the signal, whereas the energy operator reflects the variations in both amplitude and frequency of the speech signal.

The energy operator could be calculated either in the time domain or the frequency domain. In the time domain, it can be calculated using Equation (5). Whereas in the frequency domain, Equation (7) is used for the energy operator calculation.

2.3. Spectral Analysis

There are many classes of spectral analysis algorithms which are used in speech processing. The digital filter bank method is one of these algorithms that is usually used in speech processing. A filter bank can be regarded as a model of the initial transformation in the human auditory system. Three choices for the frequency axis of this bank of filters could be used in such analysis: uniform spacing (as in the standard FFT), exponential spacing (a Constant-Q or wavelet transform) or perceptually-derived spacing (the Mel or Bark scales), which is somewhere between the other choices.
The Mel scale, which is adopted in this paper and to be used with the **DEO**, is a mapping from a linear to a nonlinear frequency scale based also on human auditory perception. An approximation to the Mel-scale is:

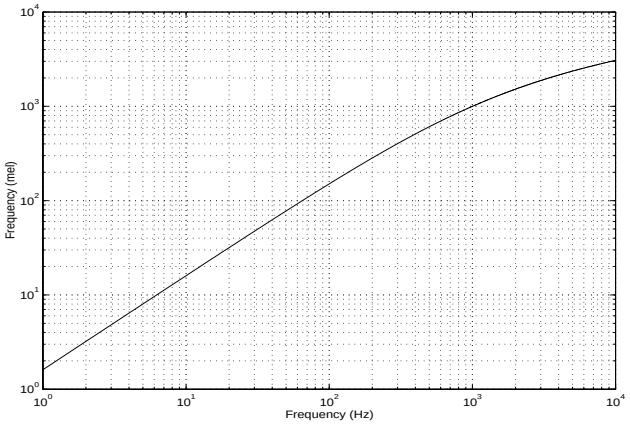

Figure 1: The Mel scale as a function of the acoustic frequency.

$$mel(f) = 2595\ log_{10}\ (1 + \frac{f}{700}), \quad (8)$$

where f corresponds to the linear frequency scale. This scale is displayed in Figure (1). It is often approximated as a linear scale from 0 to 1000 Hz, and then a logarithmic scale beyond 1000 Hz. The bandwidths of the filters used on a perceptual mel scale at a given frequency can be computed using the following transformation [8]:

$$BW_{critical} = 25 + 75\left[1 + 1.4(f/1000)^2\right]^{0.69}. \quad (9)$$

One of the easiest and most efficient ways to compute a non uniformly spaced filter bank model of the signal is to perform a Fourier transform on the signal, and sample the output at the desired frequencies. However, often the spectrum is oversampled at a finer resolution than that belonging to the nonlinear scale, and each output of the filter bank, X_{Nl}, is computed as a weighted sum of its adjacent values (i.e., kind of a spectral smoothing) as follows [7]:

$$X_{Nl} = \frac{1}{N_{FB}} \sum_{k=1}^{N_{FB}} \omega_{FB}(k)\ X(f + \delta f(f,k)), \quad (10)$$

where N_{FB} represents the number of samples used to obtain the average value, ω_{FB} represents a weighting function, and $\delta f(f,k)$ represents a function that describes the frequencies in the neighborhood of f to be used in computing the average.

2.4. Cepstral Coefficients

The cepstral coefficients are used to describe the short-term spectral envelope of a speech signal. The cepstrum is the inverse Fourier transform of the logarithm of the short-term power spectrum of the signal. By the logarithmic operation, the vocal tract transfer function and the voice source are separated. Consequently, the pulse sequence originating from the periodic voice source reappears in the cepstrum as a strong peak at the quefrency lag T_o. The advantage of using such coefficients is that they reduce the dimension of a speech spectral vector while maintaining its identity. There are two ways to obtain the cepstral coefficients: FFT cepstral and LPC cepstral coefficients.

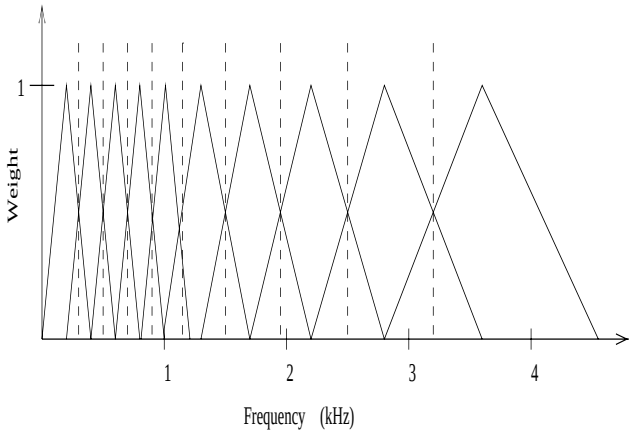

Figure 2: Filters for generating mel-frequency cepstrum coefficients.

In [6] the use of the Mel-scale (Equation (8)) in the derivation of cepstral coefficients was introduced. It was shown in this study that such a scale improves the performance of speech recognition systems over the traditional linear scale. For the **MFCC** computations, N critical bandpass filters that roughly approximate the frequency response of the basilar membrane in the cochlea of the inner ear are selected. These filters span $156 - 6844$ Hz and are spaced on the Mel-frequency scale defined in equation (8), which is roughly linear below 1 kHz and logarithmic above this frequency. The filters are triangular and multiplicatively scaled by the area. These filters are applied to the log of the magnitude spectrum of the signal which is estimated on a short-time basis. To obtain the **MFCCs**, C_n, a discrete cosine transform, is applied to the output of the N filters, X_k, as follows:

$$C_n = \sum_{k=1}^{N} X_k \; cos\Big(\frac{\pi\, n}{N}(k-0.5)\Big), n = 1, 2, ..., M, \qquad (11)$$

where M is the number of the cepstral coefficients, N is the analysis order and $X_k, k = 1, 2, ..., N$, represents the log-energy output of the k^{th} filter. For the **MFCC** computations, 20 triangular bandpass filters were simulated as shown in Figure (2).

2.5. The Hidden Markov Model Toolkit (HTK)

The speech recognition system used in our experiments, **HTK**, is completely described in [5]. **HTK** is a **HMM**-based speech recognition system. The toolkit can be used for isolated or continuous whole-word based recognition systems. The toolkit was designed to support continuous density **HMMs** with any number of state and mixture components. It also implements a general parameter tying mechanism which allows the creation of complex model topologies to suit a variety of speech recognition applications.

3. ASR USING MFCCS AND THE MDEO

3.1. Signal Representation

The feature vector, which is used to represent the speech signal in an **ASR** process, aims to preserve the information needed to determine the phonetic identity of such a signal. The best feature vector that can be used in such a process is the one that is not affected by several factors such as speaker differences, paralinguistic factors and channel effects. Given the fact that our ears are largely insensitive to phase effects, representations are almost always derived from the short-term power spectrum. In addition, this power spectrum reflects the frequency resolution of the human ear when it is based on a *mel* frequency scale. It was found in [4] that these log power spectra for speech have properties more convenient for statistically based speech recognition than the properties that are obtained by linear power spectra. Moreover, removing the correlation of energy levels in adjacent bands of the log power spectra allows the number of parameters to be reduced while preserving the useful information and consequently reduces the amount of computation needed. Given these facts, the cosine transform which converts the set of log energies to a set of largely uncorrelated cepstral **MFCCs**, when combined with the dynamic features and the energy, was found to be the most popular feature vector that could be fruitful in an **ASR** process [4].

From Equation (7) we can see that the energy operator is a function of both amplitude and frequency of the signal samples. Applying the energy operator to a speech signal to get a feature that is helpful in increasing the recognition rate of an **ASR** system came from the fact that the amplitude of the speech signal sample is always dependent on its frequency and that the traditional energy measure reflects only the amplitude of the signal, whereas the energy operator reflects the variations in both amplitude and frequency of the speech signal.

4. EXPERIMENTAL RESULTS

4.1. Database

The algorithm has been tested over a subset of both the **TIMIT** and the **NTIMIT** databases [9]. The speech was sampled at 16 kHz. The **TIMIT** corpus contains broadband recordings of 630 speakers of 8 major dialects of American English, each reading 10 phonetically rich sentences. The **NTIMIT** database is the telephone-bandwidth noisy version of the **TIMIT** database.

The baseline system used for the recognition task was a monophone/tri-phone Gaussian mixture **HMM** system. The speech was parameterized into 12 **MFCCs** along with the **MDEO** and the first differentials of these parameters. This yielded a 26-dimensional feature vector. The **TIMIT** database contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. The acoustic training data consisted of 380 sentences from the training set of both the **TIMIT** and **NTIMIT** databases in which the speech was sampled at 16 kHz. The standard **HTK** system [5] was trained using a 5-state **HMM** for each phoneme, to define 220 speech states. A single component Gaussian mixture distribution was then trained for each state, for a total of about 34320 parameters. All recognition tests were carried out on the test subset of both the **TIMIT** and **NTIMIT** databases. This test set consists of 110 sentences. The data in the **TIMIT** database was recorded in a clean environment, whereas the **NTIMIT** database is the telephone bandwidth noisy version of the **TIMIT** database.

The speech data is segmented into 25.6 msec frames with 10 msec overlapping. Each frame is weighted by a 512-point Hamming window, and then the DFT using 512-point FFT of that frame is

	$\epsilon_{Sub}(\%)$	$\epsilon_{Del}(\%)$	$\epsilon_{Ins}(\%)$	$C_{Ph}(\%)$
En	20.33	6.05	3.75	73.62
DEO	19.40	5.53	3.86	75.08
MDEO	18.46	5.63	3.65	75.91

Table 1: Recognition Performance on a subset of the **TIMIT** database for single mixture monophone using word-pair gram language model.

	$\epsilon_{Sub}(\%)$	$\epsilon_{Del}(\%)$	$\epsilon_{Ins}(\%)$	$C_{Ph}(\%)$
En	16.06	5.01	3.65	78.94
DEO	15.33	5.42	4.48	79.25
MDEO	14.70	5.63	4.38	79.67

Table 2: Recognition Performance on a subset of the **TIMIT** database for single mixture monophone using both word-pair gram language model and function-word modeling.

computed. Then the feature vector is calculated for each frame. Each vector is composed of 12 static **MFCCs**, plus the **MDEO** and dynamic coefficients. This leads to a 26-element vector per frame. To compare our new vector to the vector which uses the ordinary energy, we repeated the same calculation except that we used ordinary energy instead of using the **MDEO**.

4.2. Results for clean speech

The results of our evaluation of a subset of the entire database are listed in Tables (1) and (2). Table (1) shows the different recognition error rates for a subset of the **TIMIT** database when tests were performed using single mixture monophone acoustic models and a word-pair language model. The substitution, deletion and insertion percentage errors were defined respectively as: ϵ_{Sub}, ϵ_{Del} and ϵ_{Ins}. The average phoneme accuracy rate was represented by C_{Ph}. It is clear from this table that the average phoneme accuracy rate, C_{Ph}, is 73.62% when the traditional energy measure is included in the feature vector. However, when the **DEO** is used, instead of the traditional energy measure, the corresponding C_{Ph} is 75.08%. Moreover, the corresponding C_{Ph} is 75.91% when the **MDEO** is included in the feature vector; that is a 9.51% improvement with respect to the one that we got when we used the traditional energy measure. Table (2) shows the same tests when we used both a word-pair language model and function-word modeling. We got in this case about 3.47% improvement.

4.3. Results for telephone speech

The recognition test was performed on the telephone speech using the same standard parameter settings used for the **TIMIT** database. It was found from these tests that the inclusion of either the **DEO** or the **MDEO** with the **MFCCs** parameters instead of the traditional energy measure deteriorates the performance of the recognition process. This is due to the fact that the **DEO** is sensitive to noise if it is applied to signals which have noise added to them [1]. Consequently, we did not include these results in this paper.

5. CONCLUSION

We have proposed in this paper a new feature vector based on the mel-based discrete energy operator. Results showed that the energy calculated based on the **MDEO** performs better than the traditional energy when used combined with the **MFCC** to train the **HMMs** of an **ACSR** system for clean speech signals (**TIMIT** database). Preliminary results showed that the inclusion of such parameter in the feature vector reduces the average phoneme error rate by about 9% below than that obtained using the traditional feature vectors that are based on energy measurements. However, for telephone speech the inclusion of the traditional energy in the feature vector performs better than the proposed **MDEO**. This shows the effectiveness of adding such a parameter to the feature vector for the recognition of clean speech.

We are currently continuing the effort towards the use of other auditory-based strategies instead of the mel approximation in order to get a more robust feature vector that can be used for the recognition of both clean and telephone speech.

6. REFERENCES

[1] James F. Kaiser, "On a Simple Algorithm to Calculate The Energy of a Signal", Proc. ICASSP-90, pp. 381-384, 1990.

[2] P. Maragos, T. F. Quatieri and J. F. Kaiser, "On Amplitude and Frequency Demodulation Using Energy Operators", IEEE Trans. on Signal Processing, Vol. 41, No. 4, pp. 1532-1550, April 1993.

[3] P. Maragos, J. F. Kaiser and T. F. Quatieri, "Energy Separation in Modulations with Application to Speech Analysis", IEEE Trans. on Signal Processing, Vol. 41, No. 10, pp. 3024-3051, October 1993.

[4] L. Rabiner and B. Juang, "Fundamentals of Speech Recognition", PTR Prentice Hall Signal Processing Series, 1993.

[5] Cambridge University Speech Group, "HTK Hidden Markov Model Toolkit", Entropic Research Laboratories Inc., Cambridge, December 1993,

[6] Steven Davis and Paul Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 28, No. 4, pp. 357-366, August 1980.

[7] Joseph W. Picone, "Signal Modeling Techniques in Speech Recognition", Proceedings of the IEEE, Vol. 81, No.9, pp. 1215-1247, September 1993.

[8] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency", J. Acoust. Soc. Am. 68(5), pp. 1523-1525, November 1980.

[9] Charles Jankowski, Ashok Kalyanswamy, Sara Basson and Judith Spitz, "NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database", Proc. ICASSP-90, pp. 109-112, 1990.