Isolate Speech Recognition Based on Time-Frequency Analysis Methods

Alfredo Mantilla-Caeiros[1], Mariko Nakano Miyatake[2], and Hector Perez-Meana[2]

[1] Intituto Tecnologico de Monterrey, Campus Ciudad de Mexco,
Av. Del Puente Mexico D.F.
[2] ESIME Culhuacan, Instituto Politécnico Nacional,
Av. Santa Ana 1000, 04430 Mexico D.F. Mexico
amantill@itesm.mx, mariko@infinitum.com.mx,
hmpm@prodigy.net.mx

Abstract. A feature extraction method for isolate speech recognition is proposed, which is based on a time frequency analysis using a critical band concept similar to that performed in the inner ear model; which emulates the inner ear behavior by performing signal decomposition, similar to carried out by the basilar membrane. Evaluation results show that the proposed method performs better than other previously proposed feature extraction methods when it is used to characterize normal as well as esophageal speech signal.

Keywords: Feature extraction, inner ear model; isolate speech recognition, time-frequency analysis.

1 Introduction

The performance of any speech recognition algorithm strongly depends on the accuracy of the feature extraction method, because of that several methods have been proposed in the literature to estimate a set of parameters that allows a robust characterization of the speech signal. A widely used feature extraction method consists on applying the Fast Fourier Transform (FFT) to the speech segment under analysis. This representation in the frequency domain is obtained by using the well-known MEL scale, where the frequencies smaller than 1kHz are analyzed using a linear scale, while the frequencies larger than 1kHz are analyzed using a logarithmic scale, with the purpose of creating an analogy with the internal cochlea of the ear that works as a frequencies splitter [1]-[4].

Linear Predictive Coding (LPC) is other widely used feature extraction method whose purpose is to find set of parameters that allows an accurate representation of the speech signal as the output of an all pole digital filter, which models the vocal track, whose excitation is an impulse sequence with a period equal to the pitch period of speech signal under analysis, when the speech segment is a voiced one, or a white noise when the speech segment is an unvoiced one [1], [3]. Here, to estimate the features vector, firstly the speech signal is divided in segments of 20 to 25 ms, with 50% of overlap. Finally, the linear predictive coefficients of each segment are

E. Bayro-Corrochano and J.-O. Eklundh (Eds.): CIARP 2009, LNCS 5856, pp. 297–304, 2009.

estimated such that the mean square value of prediction error becomes a minimum. Because five formants or resonant frequencies are enough to characterize the vocal track, a predictive filter of order 10 is enough [1], [4]. Depending on the application, it may be useful to take the LPC average of the N segments contained in the word under analysis, such that this coefficients average may be used as the behavior model of a given word. Thus the averaged m-th LPC becomes

$$\hat{a}_m = \frac{1}{N}\sum_{i=1}^{N} a_{i,m}, \quad 1 \le m \le p \tag{1}$$

where N is the total number of segments contained in the word.

The cepstral coefficients estimation is other widely used feature extraction method in speech recognition problems. These coefficients form a very good features vector for the development of speech recognition algorithms [1, 2, 4], sometimes better than the LPC ones. The cepstral coefficients can be estimated from the LPC coefficients applying the following expression [1]

$$c_n = -a_n - \frac{1}{n}\sum_{i=1}^{n-1}(n-i)a_i c_{n-i} \tag{2}$$

where C_n is the n-th LPC-Cepstral coefficients, a_i is the i-th LPC coefficients and n is the Cepstral index. Usually the number of cepstral coefficients is equal to the number of LPC ones to avoid noise [1]. For isolated word recognition, it is possible to take also the average of cepstral coefficients contained in the word to generate an averaged feature vector (CLPC) to be used during the training or during the recognition task. Most widely used feature extraction methods, such as those describe above, are based on modeling the vocal tract. However if the speech signals are processed taking in account the form in which they are perceived by the human ear, similar or even better results may be obtained. Thus in [5] the use of time-frequency analysis and auditory modeling is proposed, in reference [6] an automatic speech recognition scheme using perceptual features is proposed. Thus to use an ear model-based feature extraction method may be an attractive alternative because, this approach allows characterizing the speech signal in the form that it is perceived [7].

This paper proposes a feature extraction method for speech recognition, based on an inner ear model that takes in account the fundamentals concepts of critical bands. Evaluation results using normal and esophageal speech show that the proposed approach provides better results than other previously feature extraction methods.

2 Feature Extraction Based on Inner Ear Model

In the inner ear, the basilar membrane carries out a time-frequency decomposition of the audible signal through a multi-resolution analysis similar to that performed by a wavelet transform [6]. Thus to develop a feature extraction method that emulates the basilar membrane operation, it must be able to carry out a similar decomposition, as proposed in the inner ear model developed by Zhang et. al. [8]. In this model the

dynamics of basilar membrane, which has a characteristic frequency equal to f_c, can be modeled by a gamma distribution multiplied by a pure tone of frequency f_c, that is using the so-called gamma-tone filter. Here the shape of the gamma distribution is related to the filter order, while the scale is related to the inverse of the frequency of occurrence of events under analysis, when they have a Poisson distribution. Thus the gamma-tone filter representing the impulse response of the basilar membrane is given by [8]

$$\psi_\theta^\alpha(t) = \frac{1}{(\alpha-1)!\theta^\alpha} t^{\alpha-1} e^{\frac{-t}{\theta}} \cos(2\pi t/\theta) \quad t > 0 \tag{3}$$

where α and θ are the shape and scale parameters, respectively. Equation (3) defines a family of gamma-tone filters characterized by θ and α, thus it is necessary to look for the more suitable filter bank to emulate the basilar membrane behavior. To this end, we can normalize the characteristic frequency by setting $\theta=1$ and $\alpha=3$, which according to the basilar membrane model given by Zhang et al [8], provides a fairly good approximation to the inner ear dynamics. Thus from (3) we get

$$\psi(t) = \frac{1}{2} t^2 e^{-t} \cos(2\pi t) \quad t > 0 \tag{4}$$

This function presents the expected attributes of a mother wavelet because it satisfies the admissibility condition given by [9], [10]

$$\int_{-\infty}^{\infty} |\psi(t)|^2 dt = \frac{1}{2} \int_0^{\infty} \left| t^2 e^{-t} \cos(2\pi t \right|^2 dt < \infty \tag{5}$$

That means that the norm of $\psi(t)$ in $L^2(\mathbf{R})$ space exists and then the functions given by (4) constitutes an unconditional basis for $L^2(\mathbf{R})$. This fact can be proven by using the fact that [11]

$$\int_0^{\infty} \Psi(s\omega) \frac{ds}{s} < \infty \tag{6}$$

The previous statement can verified substituting the Fourier transform of (4), Ψ(ω), into (6), where

$$\Psi(\omega) = \frac{1}{2} \left[\frac{1}{[1 + j(\omega - 2\pi)]^2} + \frac{1}{[1 + j(\omega + 2\pi)]^2} \right] \tag{7}$$

Thus we can generate the expansion coefficients of an audio signal *f(t)* by using the scalar product between *f(t)* and the function $\psi(t)$ with translation τ and scaling factor s as follows [11]

$$\gamma(\tau, s) = \frac{1}{\sqrt{s}} \int_0^{\infty} f(t)\psi\left(\frac{t-\tau}{s}\right)dt \tag{8}$$

A sampled version of (8) must be specified because we require recognizing discrete time speech signals. To this end, a sampling of the scale parameter, s, involving the psychoacoustical phenomenon known as critical bandwidths will be used [10].

The critical bands theory models the basilar membrane operation as a filter bank in which the bandwidth of each filter increases as its central frequency increases [8, 9]. This statement allows defining the Bark frequency scale; a logarithmic scale in which the frequency resolution of any section of the basilar membrane is exactly equal one Bark, regardless of its characteristic frequency. Because the Bark scale is characterized by a biological parameter, there is not an exact expression for it, given as a result several different proposals available in the literature. Among them, the statistical fitting provided by Schroeder et al [10], appears to be a suitable choice. Thus using the approach provided by [8], the relation between the linear frequency, *f,* given in Hz and the Bark frequency, *Z*, is given by [10]

$$Z = 7\ln\left(\frac{f}{650} + \sqrt{\left(\frac{f}{650}\right)^2 + 1}\right) \tag{9}$$

Next by using the expression given by (9), the central frequency in Hz corresponding to each band in the Bark frequency scale becomes [10]

$$f_c = 325 \cdot \frac{e^{\frac{2j}{7}} - 1}{e^{\frac{j}{7}}} \quad j = 1,2,\ldots \tag{10}$$

Next, using the central frequencies given by (10) the jth scale factor is given by

$$s_j = \frac{1}{f_c} = \frac{1}{325} \cdot \frac{e^{\frac{j}{7}}}{e^{\frac{2j}{7}} - 1} \quad j = 1,2,\ldots \tag{11}$$

The inclusion of bark frequency in the estimation of scaling factor, as well as the relation between (4) and the dynamics of basilar membrane, allows frequency decomposition similar to that carried out in the human hearing. The scaling factor give by (11) satisfies the Littlewood-Paley theorem since

$$\lim_{j\to+\infty} \frac{s_{j+1}}{s_j} = \lim_{j\to+\infty} \frac{e^{(j+1)/7}\left(e^{2j/7} - 1\right)}{e^{j/7}\left(e^{2(j+1)/7} - 1\right)} = \lim_{j\to+\infty} \frac{e^{(3j+1)/7}}{e^{(3j+2)/7}} = e^{-1/7} \neq 1 \tag{12}$$

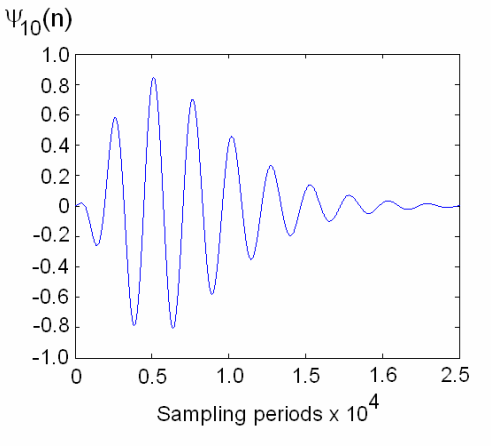

Fig. 1. 10th Gammatone function derived from the inner ear model

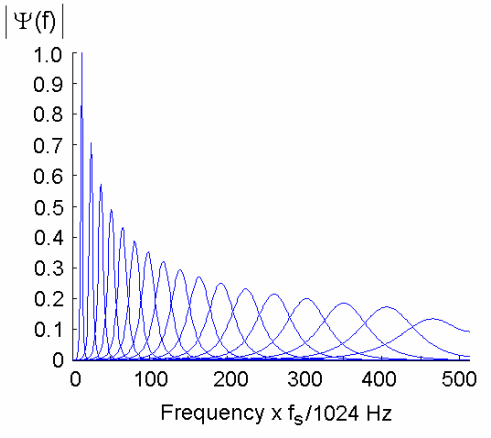

Fig. 2. Frequency response of filter bank derived from an inner ear model

Then there is not information loss during the discretization process. Finally the number of subbands is related with the sampling frequency as follows

$$j_{\max} = \text{int}\left(7\ln\left(\frac{f_s}{1300} + \sqrt{\left(\frac{f_s}{1300} \right)^2 + 1} \right) \right) \tag{13}$$

Thus for a sampling frequency equal to 8KHz the number of subbands becomes 17. Finally, the translation axis is naturally sampled because the input data is a discrete time signal, and then the expansion coefficients can be estimated as follows [9]

$$C_{f,\psi}(\tau)=\sum_{-\infty}^{\infty} f(n)\psi_s(n-\tau) \tag{14}$$

where

$$\psi_s(n)=\frac{1}{2}(nT/s)^2 e^{-(nT/s)}\cos\left(2\pi nT/s\right)) \;\; n>0 \tag{15}$$

where T denotes the sampling period. Here the expansion coefficients $C_{f,\psi}$ obtained for each subband are used to carry out the recognition task. Figures 1 shows $\psi_{10}(n)$, and Fig. 2 shows the filter bank power spectral density, respectively.

3 Evaluation Results

The performance of proposed feature extraction method was evaluated in isolate word recognition tasks, with normal as well as esophageal speech signals. Here the feature vector consists of the following parameters: the *m-th* frame energy given by [12]

$$\bar{x}_m(n)=\gamma\bar{x}_m(n-1)+x_m^2(n),\;\; n=1,2,..,N\,, \tag{16}$$

where (N=1/γ), the energy contained in each one of the 17 wavelet decomposition levels,

$$\bar{y}_{k,m}(n)=\gamma\bar{y}_{k,m}(n-1)+y_{k,m}^2(n),\quad k=1,2,...,17\,, \tag{17}$$

the difference between the energy of the previous and actual frames,

$$v_0(m)=\bar{x}_m(N)-\bar{x}_m(N-1)\,, \tag{18}$$

together with the difference between the energy contained in each one of the 17 wavelet decomposition levels of current and previous frames,

$$v_k(m)=\bar{y}_k(m)-\bar{y}_k(m-1),\quad k=1,2,...,17\,. \tag{19}$$

where m is the number frame. Then the feature vector derived using the proposed approach becomes

$$\mathbf{X}(m)=\left[\bar{x}_m(N),\bar{y}_{1,m}(N),..,\bar{y}_{17,m}(N),\bar{v}_0(m),\bar{v}_1(m),..,\bar{v}_{17}(m)\right] \tag{20}$$

Here the last eighteen members of the feature vector include the spectral dynamics of speech signal concatenating the variation from the past feature vector to the current one.

To evaluate the actual performance of proposed approach it was compared with the performance provided by others conventional methods like Mel Frequency Cepstral

Coefficients (MFCC), Linear Prediction Coefficients (LPC), Dubechies wavelet function [9] and Haar transform [9] when they are required to perform isolate work recognition tasks, using a data base developed with the assistance of Institute of Human Communication of The National Rehabilitation Institute of Mexico: The data base was developed using a 1.7GHz DELL Inspiron 8200 Pentium 4-M , with a Sony F-V220 Dynamic Microphone and an audio board Crystal WDM Audio from Cirrus Logic Inc. The data base consists of 100 words of 20 normal speakers and 20 esophageal speakers. Evaluation results provided in Table 1 shows that proposed approach provides better recognition performance than other widely used feature extraction methods. In all cases the feature vectors were estimated in similar form, with 100 words of 20 different speakers, and used as input of a recursive neural network [11]. Here half words were used for training and half for testing. Finally table 2 shows the performance of proposed approach when is used with two different pattern classification methods, the neural network and hidden Markov Model.

Table 1. Comparison between several features extraction methods using normal and esophageal speeker voice

	Proposed	Daub 4	Haar	LPC	MFCC
Normal Speaker	**97%**	**83%**	**70%**	**94%**	**95%**
Esophageal Speaker	**93%**	**77%**	**52%**	**89%**	**90%**

Table 2. Recognition performance of proposed feature extraction method when is used with two different identification algorithms

Classifier	Normal speech	Esophageal speech
Recurrent Neural Network	**95%**	**93%**
Hidden Markov Models	**92%**	**92%**

Evaluation results show that proposed algorithm performs better than other previously proposed feature extraction methods, when it is used to recognize isolated normal speech, as well as isolated esophageal speech signals.

4 Conclusions

A new feature extraction based on an inner ear model was proposed, and applied to feature extraction in isolate word recognition for normal and esophageal speech. The evaluation results performed using real speech data show that the proposed approach, based on modeling the basilar membrane, accurately extracts perceptually meaningful data required in isolate word recognition; providing better results than others feature extraction methods. The use of artificial neural network as a classifier produced

success rate higher than 97% in the recognition of Spanish word pronounced by normal speakers and 93% when the words are pronounced by esophageal speaker. An important consequence from the use of multi-resolution analysis techniques is that high frequency information is captured during the feature extraction stage.

Acknowledgements

The authors thank the National Science and Technology Council for the financial support during the realization of this research; and to Dr. Xochiquetzal Hernandez from The National Rehabilitation Institute of Mexico for the assistance provided to develop the speech database used to evaluate the proposed method.

References

1. Rabiner, L., Juang, B.: Fundamentals of Speech Recognition. Prentice Hall, Piscataway (1993)
2. Rabiner, R., Juang, B.H., Lee, C.H.: An Overview of Automatic Speech Recognition. In: Lee, C.H., Soong, F.K., Paliwal, K.K. (eds.) Automatic Speech and Speaker Recognition: Advanced Topics, pp. 1–30. Kluwer Academic Publisher, Dordrecht (1996)
3. Junqua, C., Haton, J.P.: Robustness in Automatic Speech Recognition. Kluwer Academic Publishers, Dordrecht (1996)
4. Pitton, J.W., Wang, K., Juang, B.H.: Time-frequency analysis and auditory modeling for automatic recognition od speech. Proc. of The IEEE 84(9), 1109–1215 (1999)
5. Haque, S., Togneri, R., Zaknich, A.: Perceptual features for automatic speech recognition in noise environments. Speech Communication 51(1), 58–75 (2009)
6. Suarez-Guerra, S., Oropeza-Rodriguez, J.: Introduction to Speech Recognition. In: Perez-Meana, H. (ed.) Advances in Audio and Speech Signal Processing; Technologies and Applications, pp. 325–347. Idea Group Publishing, USA (2007)
7. Childers, D.G.: Speech Processing and Synthesis Toolboxes. Wiley and Sons, New York (2000)
8. Zhang, X., Heinz, M., Bruce, I., Carney, L.: A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. Acoustical Society of America 109(2), 648–670 (2001)
9. Rao, R.M., Bopardikar, A.S.: Wavelets Transforms, Introduction to Theory and Applications. Addison Wesley, New York (1998)
10. Schroeder, M.R., et al.: Objective measure of certain speech signal degradations based on masking properties of the human auditory perception. In: Frontiers of Speech Communication Research. Academic Press, London (1979)
11. Freeman, J., et al.: Neural Networks, Algorithms, Applications and Programming Techniques. Addison-Wesley, New York (1991)
12. Mantilla-Caeiros, A., Nakano-Miyatake, M., Perez-Meana, H.: A New Wavelet Function for Audio and Speech Processing. In: Proc. of the MWSCAS 2007, pp. 101–104 (2007)