EAR-MODEL DERIVED FEATURES FOR AUTOMATIC SPEECH RECOGNITION

Renato De Mori[1], *Dario Albesano*[2], *Roberto Gemello*[2] *and Franco Mana*[2]

[1] LIA CERI-IUP
University of Avignon , BP 1228
84911 Avignon Cedex 9 - France
renato.demori@lia.univ-avignon.fr

[2] CSELT
Centro Studi e Laboratori Telecomunicazioni
via G. Reiss Romoli, 274 - 10148 Torino - Italy
dario.albesano@cselt.it, roberto.gemello@cselt.it,
franco.mana@cselt.it

ABSTRACT

The paper provides a theoretical justification that gravity centers (GC) in frequency bands computed from zero-crossing information are far more robust to additive telephone noise than GCs computed from FFT spectra. Experiments on two different corpora confirm the theoretical results when GCs are added to standard Mel Frequency-scaled Cepstral Coefficients (MFCC) and their time derivatives. A 20.1% word error reduction is observed on a large telephone corpus of Italian cities, with an average Signal-to-Noise Ratio (SNR) of 15 dB, if GCs are computed from zero-crossings, while performance deteriorates when GCs are computed from FFT spectra.

1. INTRODUCTION

An interesting pattern of research for robust acoustic features is to investigate how they are distorted by the presence of noise and how distortion can be reduced by a suitable choice of the acoustic parameters and the algorithms which use them to compute features.

For practical reasons, it is advantageous to start with simple acoustic features which can be integrated with parameters that have already been proven to be effective. Gravity centers (GC) in formant frequency bands are good candidates and will be considered in this study as parameters to be used in addition to classical Mel Frequency-scaled Cepstral Coefficients (MFCC) and their time derivatives.

The use of GCs and their trajectories in Automatic Speech Recognition (ASR) is not new [3][13]. Nevertheless, there is a number of issues which has not been addressed so far.

A first problem concerns the way GCs are computed and how this affects robustness. Simple theoretical considerations show that computation based on zero-crossing intervals at the outputs of the filters of an auditory model produced GCs which are more robust in presence of noise. Experiments confirm this conjecture showing that GCs computed with zero-crossing intervals and added to classical MFCCs and their time derivatives significantly outperform the same set of features with GCs computed from Fast Fourier Transformation (FFT).

A second problem worth to be studied in more detail, is about the recognition paradigm in which new acoustic parameters are used. The introduction of new acoustic features may have a different impact on recognition performance depending on the type of models and search process that generate hypotheses using the features as observations. Popular models are Hidden Markov Models (HMM) and connectionist models. The results of the introduction of the same set of features may strongly depend on the model type. In principle, connectionist models seem to be potentially more flexible than HMMs to the addition of new observations, but this has to be verified. If the two models have different performance, the integration of their results may lead to further improvement.

This paper discusses the use of GCs with MFCCs and their time derivatives as inputs to a Neural Network (NN) whose outputs are scores for phonemes and transitions between phonemes. The NN outputs are then used in an HMM based recognition system. The effectiveness of the added parameters is evaluated with a large vocabulary speaker-independent ASR system using a corpus of speech data collected over the telephone network.

Neural networks are useful tools for classifying phonemes or other phonetic events because they allow to integrate a variety of different features. In particular, cepstral coefficients can be integrated with features extracted directly in the spectral domain. It is also important to notice that GCs should have an impact on the output scores that depends on how reliable they are. In fact, they are less reliable when spreading is high and energy is low. Neural networks can take this into account by making high spading and low energy act as inhibitors making the influence of GCs on the outputs close to uniform. Furthermore, different degrees of inhibition are applied in different band with a beneficial effect when noise is predominant in one band only.

Section 2 of the paper discusses relevant properties of zero-crossing information when noise is added to the speech signal, Section 3 describes the ear model used for providing the acoustic parameters, including zero-crossings, with which GCs have been computed. Section 4 reports experimental results.

2. THE COMPUTATION OF GRAVITY CENTERS FROM ZERO-CROSSING INFORMATION

This section investigates the robustness of GCs computed from the zero crossings at the output of an ear model. For the sake of simplicity, the computation in a time frame is considered without explicit reference to the time reference. It is well known [2] that the zero-crossing density μ of a signal is proportional to the square-root of the ratio between the second order moment

0-7803-6293-4/00/$10.00 ©2000 IEEE.

and the zero order moment of the spectrum S(f), i.e., for a clean signal S:

$$\mu_S = k\sqrt{\frac{\sum_{f=f_1}^{f_2} f^2 S(f)}{\sum_{f=f_1}^{f_2} S(f)}} \qquad (1)$$

The zero-crossing density μ_X of a signal resulting from the perturbation of the clean signal with additive noise and having spectrum X(f) is given by:

$$\mu_X = k\sqrt{\frac{\sum_{f=f1}^{f2} f^2\{S(f)+b\}}{\sum_{f=f1}^{f2}\{S(f)+b\}}} = k\sqrt{\frac{\sum_{f=f1}^{f2} f^2 S(f)}{\sum_{f=f1}^{f2}\{S(f)+b\}} + \frac{b\sum_{f=f1}^{f2} f^2}{\sum_{f=f1}^{f2}\{S(f)+b\}}} =$$

$$= \sqrt{\mu_S^2\frac{\rho}{\rho+1} + k^2\frac{\bar{f}^2}{\rho+1}} \qquad (2)$$

The (2) shows how noise perturbs the square of the zero-crossing density. A relation, derived in [1], for GCs computed from the signal spectrum is similar to the (2) when GCs replace the square of the zero-crossing densities and the square root is removed. This suggests that the densities are more robust to noise. A similar conclusion can be reached in the interesting case in which [10], x(t) is the output of a filter and represents a noisy signal made of a speech signal harmonic having frequency $f_0 = \frac{\omega_0}{2\pi}$, plus additive noise n(t). It is possible to analyze the accuracy with which the zero-crossing density estimates the frequency of the harmonic component. Let the noisy signal be represented by the following expression:

$$x(t) = A\cos\omega_0 t + B\sin\omega_0 t + n(t) \qquad (3).$$

If A and B are independent normal random variables with the same dispersion, a bandwidth $\left\{f_1 = \frac{\omega_1}{2\pi}, f_2 = \frac{\omega_2}{2\pi}\right\}$ is considered where the spectrum of additive noise is constant w.r.t. time and frequency with power N. If σ^2 is the dispersion of A and B, then ,the autocorrelation of x(t) can be expressed as:

$R(\tau) = \sigma^2\cos\omega_0\tau + R_n(\tau)$

$R_n(\tau)$ being the autocorrelation of the additive noise.

The probability of having a zero crossing in a short interval τ • is given by:

$$p_1(\tau) = \frac{1}{\pi}\sqrt{\frac{\pi\omega_0^2\sigma^2 + N(\omega_2^3 - \omega_1^3)/3}{\pi\sigma^2 + N(\omega_2 - \omega_1)}}\,\tau = G\tau \qquad (4)$$

If μ is the zero crossing density, then: $\mu = \frac{p_1(\tau)}{\tau} = G$ which can be interpreted as twice the expected value of frequency estimation using zero-crossings in a band. As σ^2 is proportional to the signal energy, for a qualitative comparison we can assume: $\rho = q\frac{\sigma^2}{N}$, q<1.

G in the (4) can be further rewritten as follows:

$$G = \frac{1}{\pi}\sqrt{\frac{4\pi^3 f_0^2\sigma^2 + 8\pi^3 N(f_2^3 - f_1^3)/3}{\pi\sigma^2 + 2\pi N(f_2 - f_1)}} =$$

$$= 2\sqrt{\frac{f_0\sigma^2 + \frac{2}{3}N(f_2 - f_1)(f_2^2 + f_1 f_1 + f_1^2)}{\sigma^2 + 2N(f_2 - f_1)}}$$

Letting : $\tilde{f} = \frac{(f_2^2 + f_1 f_1 + f_1^2)}{3}$, $W = (f_2 - f_1)$, one gets:

$$\frac{G}{2} = \sqrt{\frac{f_0^2\frac{\sigma^2}{N} + 2W\tilde{f}}{\frac{\sigma^2}{N} + 2W}} = \sqrt{\frac{f_0^2 + \frac{2qW\tilde{f}}{\rho}}{1 + \frac{2qW}{\rho}}} \qquad (5)$$

Unfortunately, the zero-crossing density, even for a clean signal made of a damped sinusoid, has a certain variance when the signal is filtered by a narrow band filter as shown in [8]. An analysis of the variance of zero-crossing intervals is proposed in [11] and extended in [7] to show the effect of a shift in the level crossing. The case of a sine wave corrupted by white additive noise is studied in detail leading to the following expression for the variance of zero-crossing intervals:

$$\sigma_X^2 = k_\sigma \frac{1}{\rho}\frac{B}{W\omega_i^2}\frac{1}{1-\varepsilon} \qquad (6)$$

ω_i is the angular frequency of the signal, B is the noise bandwidth, W is the filter bandwidth and ε is the square of the ratio between the level crossing value and the peak amplitude of the signal.

Nevertheless, an inspection of the (5) shows that the effect of low SNR values can be compensated by a choice of W making the zero-crossing density more robust in presence of noise. Thus zero-crossing densities are expected to be noisy in a way that do not depend too much on the noise affecting the signal before clipping.

3. ARCHITECTURE FOR THE COMPUTATION OF GRAVITY CENTERS FROM ZERO-CROSSING INTERVALS

The use of zero-crossing intervals for spectral analysis has been widely investigated. An interesting overview is presented in [6] where the dominant frequency principle is stated, based on which zero-crossing analysis of a signal tends to enhance a dominant signal component. Other interesting results can be found in [5]. Zero-crossing based spectral analysis is proposed in [7], where spectral estimation is based on zero-crossing intervals extracted from the outputs of a filter bank inspired by an ear model. Zero-crossing intervals are extracted from each filter output and analyzed in a time window whose length depends on the filter in such a way that it contains on the average a number of zero-crossing intervals equal for all filters. The same architecture, named Zero Crossing and Peak Amplitude (ZCPA) has been used in the experiments described in this paper. The ZCPA model consists of a bank of bandpass cochlear filters, simulating the Basilar Membrane, each one followed by a nonlinear signal processing to simulate the transformation of the mechanical vibrations of the basilar membrane into neural firings of auditory nerve fibers, as shown in Fig. 1.

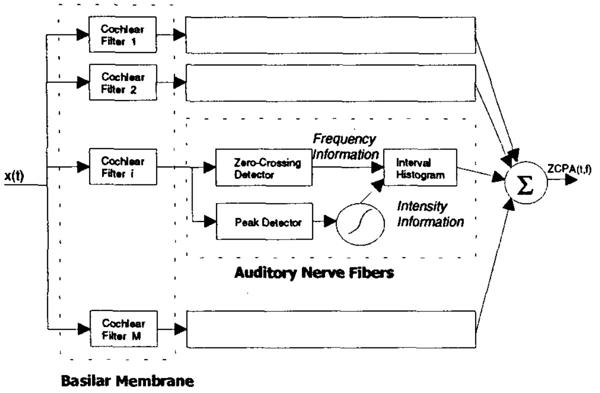

Fig. 1: Functional scheme of ZCPA Ear Model

Auditory nerve fibers firing is simulated as the upward-going zero-crossing event of the signal at the output of each bandpass filter, and the inverse of time interval between adjacent neural firings is represented as a frequency histogram, where each bin is spaced by one bark. Further, each peak amplitude between successive zero-crossings is detected, and this peak amplitude is used as intensity information. The frequency information of the signal is represented by zero-crossing intervals of sub band signals, and intensity information of the signal is given by a peak detector. The histograms across all filter channels are combined to represent the pseudo-spectrum of the auditory model, and each sub band power is compressed with a logarithmic function.

This pseudo-spectrum is represented by 16 intensity values each 10 ms corresponding to Bark spaced bands.

A comparison of spectra obtained with the zero-crossing intervals and the spectra obtained from Lineal Predictor Coefficients (LPC) show that the two spectra are very different for the same signal, the spectra obtained with zero-crossing intervals are less sensitive to noise. Furthermore, they exhibit less variance when the analysis is performed on a sine wave corrupted by additive white noise, even if in presence of a level crossing shift. Furthermore, spectra computed from zero-crossing intervals exhibit sharp peaks in the zones of energy concentration [7].

As GCs in perceptually significant bands are acoustic correlates of features like place and manner of articulation, these parameters have been widely investigated in the past for structural descriptions of speech patterns.

Zero-crossing histograms, obtained for each frame window by collecting the intervals at the output of each filter, were first proposed in [11]. In [3] a frequency interval $\{f_1,f_2\}$was subdivided into bins each containing a mass of pertinent zero-crossing intervals and center of gravity were computed for each frame and each band.

A novel computation is proposed in the following. Let z_i be the interval between two successive zero-crossings with positive slope at the output of a cochlear filter. A frequency $f_i = \frac{1}{z_i}$ and an amplitude a_i are associated to z_i. The amplitude is given by the application of the energy logarithmic function to the peak value in the interval z_i. In a given time frame, the gravity center in the band $\{f_1,f_2\}$ is obtained by considering all the contributions in that frame by z_i whose frequency falls into $\{f_1,f_2\}$. The center of gravity of the predominant harmonics of the signal in the interval $\{f_1,f_2\}$ is given by:

$$CS = \frac{\sum_{i=1}^{N_s} a_i f_i}{\sum_{i=1}^{N_s} a_i} = \frac{NS}{DS}.$$

If now noise is added to the signal, it is reasonable to assume, based on the conclusions of the previous section, that zero-crossings at the output of the filters where the harmonics dominate noise will be the ones of the signal harmonics, leading to the following computation for the gravity center of the noisy signal:

$$CX = \frac{\sum_{i=1}^{N_s} a_i f_i + \sum_{n=1}^{N_n} a_n f_n}{\sum_{i=1}^{N_s} a_i + \sum_{n=1}^{N_n} a_n} = \frac{NS + NN}{DS + DN} \qquad (7)$$

Letting:

$$CN = \frac{\sum_{n=1}^{N_n} a_n f_n}{\sum_{n=1}^{N_n} a_n} = \frac{NN}{DN}; \qquad \rho' = \frac{DS}{DS + DN}$$

one gets:

$$CX = \rho' CS + (1 - \rho')CN \qquad (8)$$

If the characteristics of the ear model tend to emphasize signal harmonics, then DN should be much lower than what it would be without the signal taking ρ' close to 1. Obviously this is not the case when the signal is not made of sufficiently strong harmonics. For this reason, the *spreading* of the contributions to the summation around the gravity center is used to take into account the reliability of the feature.

4. EXPERIMENTAL RESULTS AND CONCLUSIONS

Experiments were conducted using a hybrid system described in [4] with a feed-forward Neural Network (NN) which computes the probability of being in a state of a Hidden Markov Model (HMM) given the observation made of a set of input frames. Separate train and test corpora were used for the experiments described below.

Both corpora are made of telephone speech in the 300-3400 Hz band sampled at 8 kHz. A total of 1136 speakers provided speech samples.

Speakers were evenly distributed among males and females coming from many Italian regions and with different accents. Training was performed on 1136 speakers uttering a total of 4875 phonetically balanced sentences with a vocabulary of 3653 words.

Two tests with the same corpus were performed. The corpus consists of 14473 isolated word utterances, from 1050 speakers, containing 475 city names.

Training and test corpora were collected from telephone calls from different locations all over Italy. An additional test set was derived from the first one by adding real telephone background noise with a resulting Signal-to-Noise Ratio (SNR) of 15 dB.

The first recognition test used a vocabulary of 9329 names of Italian towns and villages, thus with a perplexity of 9329. Two experiments were performed, one indicated as T1C with the original test set and another one, indicated as T1N, with the corrupted signal.

Table 1 reports the experimental results for T1C and T1N, Rows indicate the set of parameters provided at the input of NN. Base indicates the usual set of 12 MFCC plus total signal energy, with their first and second derivatives. GCFFT indicates the set of three GCs computed from FFT spectra in the following frequency bands (with a triangular weighting) roughly corresponding to the bands of the first three formants:

B1 = {0 , 1175} Hz,
B2 = {315,2860} Hz,
B3 = {1175,4000} Hz.

GCOX indicate the set of GCs in the same bands, but computed with zero-crossing information.

IS indicates that the importance of GCs is modulated by applying at the NN inputs the energy intensity as well as the spreading of energy in each band, the latter is computed as the percentage of the band occupied by 80% of the energy in the band.

In [9] it is reported that the use of spectral sub-band GCs in addition to 10 MFCC leads to an improvement in the recognition of spoken letters belonging to the e-set.

	T1C	T1N
Base	**83.45**	**51.85**
Base + GCFFT	**85.15**	**47.97**
Base + GCFFT+IS	**86.76**	**43.02**
Base + GCOX+IS	**87.11**	**56.31**

Table 1. Recognition accuracy with 9329 city names

The same tests were performed with a closed vocabulary of 475 city names. The results are reported in Table 2 where T2C refers to the original signals and T2N refers to the noisy ones.

	T2C	T2N
Base	**94.22**	**72.63**
Base + GCFFT	**94.87**	**69.56**
Base + GCFFT + IS	**95.38**	**66.36**
Base + GCOX + IS	**95.48**	**78.13**

Table 2. Recognition accuracy with 475 city names

REFERENCES

[1] D. Albesano, R. De Mori, R. Gemello, F. Mana, "A Study on the Effect of Adding New Dimensions to Trajectories in the Acoustic Space", in *Proc. of Eurospeech'99,* Budapest, Hungary, 1999, pp. 1503-1506.

[2] A.Chang, Phil, Essigman, "Representation of speech sound and some of their acoustical properties". *Proc. of the Institute of Radio Engineers*, vol. 39, 1951

[3] R. De Mori, "A Descriptive Technique for Automatic Speech Recognition", *IEEE Transactions on Audio and Electroacoustics*, Vol. Au-21, No. 2, 1973, pp. 89-100.

[4] R. Gemello, D. Albesano, F. Mana "Continuous Speech Recognition with Neural Networks and Stationary-Transitional Acoustic Units", in *Proc. of IEEE International Conference on Neural Networks (ICNN-97)*, Houston, USA 1997, pp.2107-2111.

[5] S.M. Kay and R. Sudhaker. "A zero-crossing based spectrum analyzer", *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-34(1):96-104, 1986.

[6] B. Kedem, "Spectral analysis and discrimination by zero-crossings", *Proceedings of thr IEEE*, 74(11):1477-1493., 1986.

[7] D.S. Kim, S.Y. Lee. and R.M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environment", *IEEE Transactions on Speech and Audio Processing*, SAP-7(1): 55-69,1999.

[8] R.J. Niederjohn and M. Lahat, "A zero-crossing consistency method for formant tracking of voiced speech in high noise level", *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-34(2):349-355, 1985.

[9] K.K. Paliwal, "Spectral Subband Centroid Features for Speech Recognition", *Proc. of International Conference on Audio, Speech and Signal Processing*, 1998,p.617-620.

[10] A. Papoulis, "Probability, Random variables and Stochastic processes", Mc Grow Hill, 1965.

[11] T. Sakai T. and Doshita, "The automatic speech recognition system for conversational sound", *IEEE Transactions on Electronic Computers*, EC12:835-846, Dec. 1963.

[12] T.V. Sreenivas and R.J. Niederjohn, "Zero-crossing based spectral analysis and SVD spectral analysis for formant frequency estimation in noise", *IEEE Transactions on Signal Processing*, SP-40(2):282-293, 1992.

[13] D.X. Sun, "Robust estimation of spectral center-of-gravity trajectories using mixture spline models", *Proc. of Eurospeech'95*, 749-752, 1995.