A STUDY ON THE EFFECT OF ADDING NEW DIMENSIONS TO
TRAJECTORIES IN THE ACOUSTIC SPACE

D. Albesano§, R. De Mori(*), R. Gemello§, F. Mana§

(*) University of Avignon
§ CSELT – Centro Studi e Laboratori Telecomunicazioni
dario.albesano@cselt.it, renato.demori@lia.univ-avignon.fr,
roberto.gemello@cselt.it, franco.mana@cselt.it

ABSTRACT

The paper discusses the use, in a hybrid recognizer, of gravity centers (gc) in spectral subbands as features to be used in addition to Mel Scaled Cepstral Coefficients (MFCC) and their time derivatives. Results on noisy telephone speech show that gc computed after the nonlinear processing of an ear model increase the word accuracy from 72.63% to 78.13%.

1. INTRODUCTION

The success of most existing systems for Automatic Speech Recognition (ASR) rely on the fact that a set of well chosen acoustic features lead to interesting performance when used in a suitable, well defined modeling and recognition framework. Nevertheless, when systems are deployed outside research laboratories, their limitations appear evident and areas of possible improvement have been identified.

One of them is the search for robust features. An interesting pattern of research along this line is to investigate acoustic features in terms of how much they are distorted by the presence of noise and how distortion can be reduced by suitable enhancing algorithms. 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.

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

2. GRAVITY CENTERS

Gravity centers in broad frequency bands are indicators of frequencies of great energy concentrations. These frequencies are expected to be related to vocal tract resonances and their computation is much simpler and robust than formant tracking.

The use of gravity centers and their trajectories in Automatic Speech Recognition (ASR) is not new (see De Mori, [1]). Recently Paliwal [4] proposed the use of spectral subband gravity centers as features to be used as supplement to cepstral features for speech recognition. Gravity centers in spectral subbands are computed by splitting the power spectrum into frequency subbands $F_m$, ranging from a lower edge $l_m$ to a higher edge $h_m$ and applying the classical definition of gravity center.

For subband $F_m$, the gravity center is defined as:

$$C_m(f) = \frac{\int_{l_m}^{h_m} f w_m(f) X(f) df}{\int_{l_m}^{h_m} w_m(f) X(f) df} \quad (1)$$
where \( f \) is frequency, \( X(f) \) is the power spectrum and \( w_m(f) \) is a weighting function that filters the power spectrum in the subband which, among other forms, can be rectangular or triangular.

Figure 1 shows the time evolution of three gravity centers for a utterance of the Italian word Genova (jh eh n ow v ah. Time is measured in msecs. and frequency in kHz. The gravity centers have been computed in subbands roughly corresponding to the bands of the first three formant frequencies. Short-time spectra are computed with Fast Fourier Transform (FFT).

The time evolution of the overall signal energy is also represented in the figure by a thicker line whose main peaks correspond to the three vowels in the uttered word. This curve also shows the presence of noise peaks especially before the utterance.

The three gravity centers are computed inside three subbands positioned on 0-1175 Hz, 315-2860 Hz, 1175-4000 Hz, weighted with a triangular filter.

Fig. 2 shows gravity centers C1, C2, C3 corresponding to the utterance of the sequence of the five Italian vowels “a-e-i-o-u” (ah, eh, iy, ow, uh).

Fig. 3 shows the trajectory described by the same gravity centers in the space C1, C2, C3. A zone in the middle can be noticed where noise is concentrated. Five well separated zones can also be observed, corresponding to the five vowels, indicating that gravity centers are potentially good discriminative features for vowels uttered in noise.

3. EFFECT OF NOISE ON GRAVITY CENTERS

Unfortunately, gravity centers are badly affected by noise, as it is shown by the following simple considerations.

In general, assuming convolutive and additive noise, the spectrum of a speech frame can be represented as follows:

\[
X(f) = H(f)S(f) + N(f)
\]

where \( X(f) \) is the spectrum of the noisy signal, \( H(f) \) is the spectrum of the frequency response of the
channel, $S(f)$ is the speech spectrum and $N(f)$ is the spectrum of the additive noise.

As convolutional noise is present in the training as well as in the test corpus, the main purpose is that of reducing the effects of the additive noise which causes the mismatch between train and test conditions. In this framework, the clean signal $S(f,t)$ is the one that corresponds to the training conditions. Furthermore, it is reasonable to assume that the additive noise is constant in time intervals between two successive intervals of absence of speech signal in which the noise spectral samples can be acquired. Under these conditions the time-varying spectrum $X(f,t)$ of the observed signal $x(t)$ can be expressed as follows:

$$X(f,t) = S(f,t) + N(f)$$  \hspace{1cm} (2)

Using, for the sake of simplicity, a rectangular weighting function, the gravity centers of the clean (CC) and noisy (CN) speech, in a given frequency band $\{f_1, f_2\}$ are obtained as follows:

$$CC(t) = \frac{\sum_{f=f_1}^{f_2} S(f,t)}{\sum_{f=f_1}^{f_2} \sum S(f,t)}$$  \hspace{1cm} (3)

$$CN(t) = \frac{\sum_{f=f_1}^{f_2} \{S(f,t) + N(f)\}}{\sum_{f=f_1}^{f_2} \sum \{S(f,t) + N(f)\}}$$

Let us introduce:

$$\gamma(t) = \frac{\sum_{f=f_1}^{f_2} X(f,t)}{\sum_{f=f_1}^{f_2} \sum N(f)}$$  \hspace{1cm} (4)

with which the (3) can be rewritten as follows:

$$CN(t) = \frac{\sum_{f=f_1}^{f_2} f S(f,t)}{\sum_{f=f_1}^{f_2} \sum S(f,t)} + \frac{\sum_{f=f_1}^{f_2} f N(f)}{\sum_{f=f_1}^{f_2} \sum N(f)}$$

leading to:

$$CN(t) = \frac{CC(t)}{\gamma(t)} \{\gamma(t) - 1\} + CM \frac{CM}{\gamma(t)}$$  \hspace{1cm} (5)

where:

$$CM = \frac{\sum_{f=f_1}^{f_2} f N(f)}{\sum_{f=f_1}^{f_2} \sum N(f)}$$

4. INCREASING ROBUSTNESS

There are essentially three dimensions along which robustness of an ASR system using gravity centers can be improved.

The first dimension consists in introducing suitable transformations which reduce the difference between $CN(t)$ and $CC(t)$. Work along this line is in progress and will be reported in a future paper.

A second possibility consists in computing gravity centers from signal transformations, like the ones of an ear model, which enhance the speech signal component with respect to degradation due to the noise component. A theoretical discussion on how these models reduce the effects of noise distortions on gravity centers is beyond the space allowed for this paper. Nevertheless, an effective computation of gravity centers from the outputs of an ear model will be presented, together with experimental results, in the next section.

A third possibility, described in the following, consists in simply introducing indicators of the reliability with which gravity centers are computed. These parameters can then be used to attenuate the importance of gravity centers in the generation of phoneme hypotheses when gravity center computation is expected to be less reliable. Two parameters are used for this purpose, namely intensity and spreading.

The gravity center intensity is defined as:

$$I_m = \int_{f_1}^{f_2} w_m(f) X(f) df$$  \hspace{1cm} (6)

and indicates the amount of energy in each subband.

The gravity center spreading is defined as:

$$S_m = \sqrt{C_m(f^2) - C_m(f)^2}$$  \hspace{1cm} (7)
A small spreading value indicates a strong energy concentration in a subband. This is often a cue for a speech sound, especially when the signal energy has a peak and the noise is spread in the subband. A spreading indicator can then modify the importance a classifier gives to spectral distributions with great variance.

Gravity centers have also been computed, in the same bands as before, using a pseudo-spectrum obtained with a Ear Model [3] which is based on cochlear bandpass filters with nonlinear processing at the output of each filter. It simulates the frequency selectivity of the basilar membrane and consists in using a peak detector and a compressive nonlinearity. A pseudo-spectrum is obtained by adding in frequency bins, indexed by zero-crossing intervals computed at the output of each filter, energy contributions provided by the nonlinear processing. Once obtained a pseudo-spectrum, the gravity centers are computed according to (1), (2) and (3).

5. EXPERIMENTAL RESULTS AND CONCLUSIONS

An hybrid HMM-NN approach, previously described in [2] has been used.

A feed-forward network has an input window that spans some contiguous frames and estimates probabilities $P(Q|X)$ of being in state $Q$ given the input observations $X$.

The HMM-NN model was trained with a corpus having the following characteristics:
- telephone quality, read speech in the 300-3400 Hz band, sampled at 8KHz;
- 1136 speakers, evenly distributed among males and females, from many Italian regions, with different accents;
- 4875 sentences with 3653 phonetically balanced words.

Training and test corpora were collected from real calls all over Italy. A first test set denoted Clean is made up by 14473 isolated word utterances from a vocabulary of 475 Italian towns. A second test set denoted Noisy is built from the first one by adding telephone box noise to obtain a 15 dB SNR. Table 1 shows the recognition results in terms of word accuracy (WA) obtained with standard features (MFCC + Energy + their first and second derivatives) and with the addition of gravity centers (gc) computed from the FFT power spectrum and from a ear model derived spectrum.

<table>
<thead>
<tr>
<th>Employed Features</th>
<th>Clean</th>
<th>Noisy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC + E + d, dd (base)</td>
<td>94.22</td>
<td>72.63</td>
</tr>
<tr>
<td>base + gc (from fft)</td>
<td>94.87</td>
<td>69.56</td>
</tr>
<tr>
<td>base + gc (from fft) + intensity and spreading</td>
<td>95.38</td>
<td>66.36</td>
</tr>
<tr>
<td>base + gc (from fft) band-pass filtered + intensity and spreading</td>
<td>95.25</td>
<td>72.07</td>
</tr>
<tr>
<td>base + gc (from ear model)</td>
<td>94.74</td>
<td>76.66</td>
</tr>
<tr>
<td>base + gc (from ear model) + intensity and spreading</td>
<td>95.48</td>
<td>78.13</td>
</tr>
</tbody>
</table>

Table 1. Recognition results with gravity centers added to MFCC.

The results shows that the addition of gc leads to a limited improvements on the clean test set (12.2 % error reduction), while the extension of gc with the intensity and spreading features leads to 20.1 % error reduction. The gc computed from FFT spectra do not appear to be very robust in the presence of additive noise, while ear model derived gc appear to be much more robust to noise.

6. REFERENCES