Statistical Discrimination and Identification of Some Acoustic Sounds
H. Sayoud*, S. Ouamour**
USTHB, Electronics Institute, BP 32 Bab-Ezzouar, Alger, Algeria
Email: * sayoud@hotpop.com ** ouamour@hotpop.com

Abstract — Given that most of the speech signal recordings are generally mixed with other sounds like music, songs, or noises and knowing that the processing of any speech signal will be easier when we separate the speech area from the non-speech area, we propose a pre-processing method for speech/ non speech discrimination which is also able to identify some acoustic sounds, by using some statistical observations (mean, standard deviation) linked to a statistic measure of similarity (µGc).

Since it has been possible to discriminate between speakers thanks to the small within-variability and the large between-variability of the speaker’s acoustic features, we thought to extend this property for the purpose of acoustic sounds discrimination. Thus, we led an investigation on different types of sounds as: noises, music and speech (speech signals are extracted from TIMIT database). The purpose of this investigation is to try to define a separate class for each type of sound according to the similarity measure µGc.

Experiments showed that the similarity distance range, between speech and other acoustic signals, has a mean and standard deviation which are specific for each sound. So, for instance it will be possible to state whether a particular audio signal is really speech or non-speech, only by observing the statistical range of the µGc which is chosen as a similarity distance.

For instance, we have deduced that thanks to the value of µGc it is possible to know if an audio frame is a pure speech or non-speech, only by observing the statistical range of the µGc which is chosen as a similarity distance.

Index Terms — speech / non-speech discrimination, statistical distances, noises, µGc.

I. INTRODUCTION

In speech recognition and speaker tracking, we often deal with the problem of speech / non-speech detection. So, it is important to check whether the considered audio frame is really speech or not.

This preliminary step is important since it reduces the errors of transcription, segmentation and tracking. Thus, without having such information about the audio signal processed, it would be very difficult and inaccurate to proceed with the necessary segmentation during the indexing process [1] [2].

In order to solve this problem, we undertook a series of experiment related to the classification of the most common available sounds / noises likely to be present during interviews (TV broadcast news, TV debates, seminars and so on) and may be added to the speech signal.

In fact, the overall idea of this research work is to try to find a pre-processing technique which allows to discriminate between the different sounds. This technique is based on the similarity measure µGc introduced by Bimbot and used for speaker recognition tasks [3].

This paper is organized as follows: Chapter 2 describes the database treated, chapter 3 gives the theory of the statistical measure of similarity. In chapter 4 we describe the technique of discrimination and in chapter 5 we show and interpret the different results obtained, finally, we give a conclusion and some references related to this research field.

II. DATABASE

We used :
- 10 pieces of music from different compositors.
- And 3 types of noise, which are:
  - The human noise representing the different speech sounds produced by human beings, such as: cough, sneeze, ...
  - The office noise, such as that provoked by chairs moving, pens fall, ...
  - The background noise, issued from TV or radio in case of receiving a random frequency (absence of channels).

From these sounds, we made 703 couples of (speaker, speaker), 40 couples of (speaker, song), 40 couples of (speaker, music) and 12 couples of (speaker, noise).
III. "SOSM" APPROACH (SECOND ORDER STATISTICAL MEASURES)

Our speaker identification method [5], based on mono-Gaussian model [3], uses some measures of similarity, which are called Second Order Statistical Measures (SOSM). These measures are used in order to recognise the speaker at each segment of the speech signal.

We recall below the most important properties of the approach [3].

Let \( \{x_i\} \) be a sequence of \( M \) vectors resulting from the \( P \)-dimensional acoustic analysis of a speech signal uttered by speaker \( x \). These vectors are summarised by the mean vector \( \bar{x} \) and the covariance matrix \( X \):

\[
x = \frac{1}{M} \sum_{i=1}^{M} x_i
\]

and

\[
X = \frac{1}{M} \sum_{i=1}^{M} (x_i - \bar{x})(x_i - \bar{x})^T
\]

Similarly, for a speech signal uttered by speaker \( y \), a sequence of \( N \) vectors \( \{y_t\} \) can be extracted.

By assuming that all acoustic vectors extracted from the speech signal uttered by speaker \( x \) are distributed like a Gaussian function, the likelihood of a single vector \( y_t \) uttered by speaker \( y \) is

\[
G(y_t / x) = \frac{1}{(2\pi)^{P/2} |det X|^{1/2}} e^{-(y_t - \bar{x})^T X^{-1} (y_t - \bar{x})}
\]

If we assume that all vectors \( y_t \) are independent observations, the average log-likelihood of \( \{y_t\} \) can be written as

\[
\bar{L}_x(y_t) = \frac{1}{N} \log G(y_1, \ldots, y_N / x) = \frac{1}{N} \sum_{t=1}^{N} \log G(y_t / x)
\]

We also define the minus-log-likelihood \( \mu(x, y) \) which is equivalent to similarity measure between vector \( y_t \) (uttered by \( y \)) and the model of speaker \( x \), so that

\[
\text{Arg max}_x \ G(y_t / x) = \text{Arg min}_x \ \mu(x, y)
\]

We have then:

\[
\mu(x, y) = -\log \ G(y_t / x)
\]

The similarity measure between test utterance \( \{y_t\} \) of speaker \( y \) and the model of speaker \( x \) is then

\[
\mu(x, y) = \mu(x, y^N) = \frac{1}{N} \sum_{t=1}^{N} \mu(x, y_t)
\]

\[
= -\bar{L}_x(y_t)
\]

After simplifications, we obtain

\[
\mu(x, y) = \frac{1}{P} \left[ -\log \left( \frac{|det(Y)|}{|det(X)|} \right) + \text{tr}(YX^{-1}) + (\bar{y} - \bar{x})^T X^{-1} (\bar{y} - \bar{x}) \right] - 1
\]

This measure is equivalent to the standard Gaussian likelihood measure (asymmetric \( \mu_{0.5} \)) defined in [3] [6].

A variant of this measure called \( \mu_{GC} \) is deduced from the previous one by assuming that \( \bar{y} = \bar{x} \) (i.e. the inter-speaker variability of the mean vector is negligible).

Thus, the new formula becomes:

\[
\mu_{GC}(x, y) = \frac{1}{P} \left[ -\log \left( \frac{|det(Y)|}{|det(X)|} \right) + \text{tr}(YX^{-1}) \right] - 1
\]

A symmetric measure can be constructed by combining \( \mu_{GC}(x, y) \) with its dual term \( \mu_{GC}(y, x) \), leading to \( \mu_{GC0.5}(x, y) \) (see formula 11).

\[
\mu_{GC0.5}(x, y) = \frac{\mu_{GC}(x, y) + \mu_{GC}(y, x)}{2}
\]

These two measures (\( \mu_{GC0.5} \) and \( \mu_{G} \)) are used in our experiments. Thus, a small distance (measure) between two utterances indicates that these utterances belong probably to the same speaker, and vice versa.

IV. SPEECH-SOUND DISTANCE AND SOUNDS CLASSIFICATION

In case of sounds classification, measuring the similarity rate by the \( \mu_{GC0.5} \) permits to have an idea on the type of considered sound.

If \( \mu_{GC0.5} \in [\text{Threshold}_{min}, \text{Threshold}_{max}] \) then we identify the considered sound as type « j ».

Thresholds \( \text{Threshold}_{min} \) and \( \text{Threshold}_{max} \) can only be estimated empirically by formulas 12 and 13.

\[
\text{Threshold}_{min} = \min(\mu_{GC0.5}(\text{speech, sound}_j))
\]

\[
\text{Threshold}_{max} = \max(\mu_{GC0.5}(\text{speech, sound}_j))
\]

Where the word « sound » represents a set of sounds with the same type and « speech » represents a set of reference speakers. \( j \) denotes a certain sound belonging to a certain class "j".
V. RESULTS AND INTERPRETATION

Experimental tests of discrimination use 703 couples of (speaker, speaker), 40 couples of (speaker, song), 40 couples of (speaker, music) and 12 couples of (speaker, noise). The corresponding results are exposed in tables 1 and 2.

Results in table I represent the distances between the speech and the different sounds (speech, speech), (speech, song) and (speech, music) with their corresponding statistics. We notice that the range of $\mu_{Gc_{0.5}}$ measure for the intra-speaker speech varies from 0.13 to 0.27, whereas for the inter-speaker speech this range varies from 0.27 to 1.8. Also, for the songs this range varies from 1.4 to 3.37 and for the music it varies from 2.52 to 4.92.

Therefore, we remark that we can make a rough discrimination between the speech signal (for a known speaker) and another sound. For instance, the $\mu_{Gc}$ range between speech and music is given by the following:

- $\mu_{Gc} < 0.27$ for a pure speech.
- $\mu_{Gc} \in [0.27, 1.8]$ for a speech of a different speaker.
- $\mu_{Gc} \in [2.52, 4.92]$ for the music.

Thus, it appears possible to roughly check whether an audio signal is a pure speech or music.

\begin{table}[h]
\centering
\caption{Distances between speech and the different sounds.}
\begin{tabular}{|c|c|c|c|c|}
\hline
 & (speech, music) & (speech, song) & (speech, speech) & (speech, speech) \\
 & Intra-speaker & & & Inter-speaker \\
\hline
Mean($\mu_{Gc}$) & 3,57 & 2,49 & 0,18 & 0,79 \\
\hline
Standard deviation($\mu_{Gc}$) & 0,56 & 0,83 & 0,03 & 0,36 \\
\hline
Min($\mu_{Gc}$) & 2,52 & 1,4 & 0,13 & 0,27 \\
\hline
Max($\mu_{Gc}$) & 4,92 & 3,37 & 0,27 & 1,8 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Distances between speech and the different noises.}
\begin{tabular}{|c|c|c|c|}
\hline
 & (speech, background noise) & (speech, office noise) & (speech, human noise) \\
\hline
Mean($\mu_{Gc}$) & 9,5 & 2,1 & 1,12 \\
\hline
Standard deviation ($\mu_{Gc}$) & 0,87 & 0,3 & 0,45 \\
\hline
Min($\mu_{Gc}$) & 8,23 & 1,67 & 0,78 \\
\hline
Max($\mu_{Gc}$) & 10,34 & 2,46 & 1,88 \\
\hline
\end{tabular}
\end{table}

Results in table II represent the distances between speech and three types of noise (human noise, office noise and background noise) with their corresponding statistics.

We notice that the $\mu_{Gc}$ range for the human noise varies from 0.78 to 1.88; the office noise range varies from 1.67 to 2.64 and the background noise one varies from 8.23 to 10.34 and knowing that the $\mu_{Gc_{0.5}}$ range for the intra-speaker speech varies from 0.13 to 0.27. We can make a preliminary discrimination between speech and certain noises: see below.

- $\mu_{Gc} < 0.27$ for a pure speech.
- $\mu_{Gc} \in [0.27, 0.78]$ for a speech of a different speaker.
- $\mu_{Gc} \in [0.78, 1.8]$ for human noise or for a speech of a different speaker.
- $\mu_{Gc} \in [1.8, 2.46]$ for office noise.
- $\mu_{Gc} \in [8.23, 10.34]$ for background noise.

VI. CONCLUSION

In the field of speech and speaker recognition, we usually try to eliminate the non-speech part from the speech signal in order to improve the recognition precision. But in practice, the speech signal is often mixed with other acoustic sounds like noises, music or even songs and knowing that the task of speech processing becomes easier when we identify and separate the speech area from the non-speech one, we have proposed a pre-processing method for speech/ non speech discrimination which is also able to identify some acoustic sounds, by using some statistical observations linked to some similarity measures as the $\mu_{Gc}$.

This method has two aims: firstly, the discrimination between speech and other sounds and secondly the classification of the different acoustical sounds according to the $\mu_{Gc}$ values, used as a measure of similarity, corresponding for each sound.

And since it has been possible to discriminate
between speakers thanks to the small within-variability and the large between-variability of the speaker’s acoustic features, we thought to extend this property / idea to the purpose of acoustic sounds discrimination.

So, in this research work, we have undertaken an investigation on the classification of some different sounds likely to be met during interviews, debates or TV news by using certain similarity measures: the SOSM. These different measures are calculated in relation to a speech reference by using the $\mu_Gc$ distance.

This investigation has allowed to discriminate and identify certain types of sounds and noises: see figure 1 below. This figure resumes the different possible classifications and the most probable sounds regarding to the $\mu_Gc$ measure range calculated between the speech reference and the analysed sound. For instance, if $\mu_Gc$ is within [2.52 – 4.92] then we can state that the considered sound should be a music (table III).

Experiments showed that the similarity distance range, between the speech reference and the other acoustic signals, has a mean and standard deviation which are specific to each type. So, for instance it will be possible to state whether a particular audio signal is really a speech or non-speech, only by observing the statistical range of the $\mu_Gc$. However, we noticed some cases of confusion when the sound corresponds to a mixture of speech and other acoustic sounds, such as songs.

Anyway, even if few sounds are roughly discriminated due to the overlapping of the $\mu_Gc$ ranges, this statistical measure of similarity has shown an interesting discrimination ability in the average. So, this pre-processing method could give good results of discrimination, especially in the absence of overlapping sounds such as songs or human noises which may cause some confusion. In such cases (no overlapping), we speak about sounds accurately identified (e.g. discrimination between speech and music, or discrimination between speech and office noise).

---

**Fig. 1.** Graphic representation of the different classes of sounds according to the $\mu_Gc$ distance.

**TABLE III : Classification of sounds according to the $\mu_Gc$ range.**

<table>
<thead>
<tr>
<th>$\mu_Gc$ range</th>
<th>probable classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.13 – 0.27</td>
<td>Speech of a known speaker</td>
</tr>
<tr>
<td>0.27 – 0.78</td>
<td>Speech of a different speaker</td>
</tr>
<tr>
<td>0.78 – 1.8</td>
<td>Speech of a different speaker or human noise</td>
</tr>
<tr>
<td>1.8 – 2.46</td>
<td>Office noise</td>
</tr>
<tr>
<td>2.52 – 4.92</td>
<td>Music</td>
</tr>
<tr>
<td>8.23 – 10.34</td>
<td>Background noise</td>
</tr>
</tbody>
</table>
REFERENCES


