PENS: a confidence parameter estimating the number of speakers

Siham Ouamour[1], Mhania Guerti[2] and Halim Sayoud[1]
[1]USTHB, Institut d'Electronique, BP 32 Bab Ezzouar, Algeria
[2]Ecole Nationale Polytechnique, Algeria

https://doi.org/10.36505/ExLing-2008/02/0045/000104

Abstract

Is it possible to know how many speakers are speaking simultaneously in case of speech overlap? If the human brain, creation not yet mastered, manages to do it and even to understand the mixed speech meaning, it is not yet the case for the existing automatic systems. For this task, we propose a new method able to estimate the number of speakers in a mixture of speech signals. The algorithm developed here is based on the computation of the statistical characteristic of the 7th Mel coefficient extracted by spectral analysis from the speech signal. This algorithm using a confidence parameter, which we called PENS, is tested on seven different sets of the ORATOR database, which contain seven multi-speaker files each. Results show that PENS parameter permits us to make a good discrimination, without any ambiguity, between a mono-speaker signal (only one speaker is speaking) and a mixed-speakers signal (several speakers are speaking simultaneously). Moreover, it permits us to estimate, in case of mixed speech signals, the number of speakers with a good precision, especially when the number of speakers is less than four.

Introduction

Often during discussions, debates and confrontations, when several speakers share a discussion, we are in presence of simultaneous speech mixture of several speakers, due to the intervention of these speakers in the same time, during the discussion: Takayuki Arai (Arai 2003). Thus, the speech signal will contain some zones of speech overlap: F. Asano (Asano 2007).

Such cases often arise with female speakers: women have a multi-task behavior (Changingminds.org) which permits them to speak and understand in such conditions, although that case may also arise with male speakers, often during hot debates between adversary presenting opposite ideas, such as political debates for example. Moreover, those speech overlaps may characterize specifically one language more than another: for instance, in certain regions of Italy (Changingminds.org) people are known by the fact to begin to speak even before the other interlocutor has finished his sentence.

However, in audio document indexing by speakers, those overlap zones remain difficult to index, since we cannot attribute them to a single speaker alone. So it is interesting to know these zones locations even before applying the indexing system. For that reason, we have developed a new algorithm able to discriminate between a mono speaker speech signal and multi-

speaker speech signal containing several speakers speaking in the same time. This algorithm has many applications: it can be applied, for instance, to an audio document, just before the indexing phase in order to avoid and eliminate the segments presenting such ambiguities.

In the rest of the paper, we will present our new algorithm and show the experimental results. We will conclude at the end of the paper by giving some discussions on the results.

Approach description

This research work deals with the estimation of the number of speakers in a speech mixture: Takayuki Arai (Arai 2003). Our approach is based on the statistical characteristics of the 7th Mel filter (mel_7): H. S. Lee and A.C. TSOI (Lee 1995), as described in table 1.

Table 1: Spectral characteristics of the 7th Mel filter

	Cut-off frequency at 0%	Cut-off frequency at 50%
Fmin	1.3750 kHz	1.6125 kHz
Fmax	2.3750 kHz	2.0813 kHz
Median Freq.	1.8375 kHz	1.8375 kHz

In reality, the discovery of a confidence parameter, estimating the number of voices in a speech signal, was found after several experimental trials, but none of the other tested parameters was interesting: only one was pertinent. We called this pertinent parameter: 'Parameter Estimating the Number of Speakers' (PENS). This parameter is given by:

$$PENS = \overline{mel}_7 - \sqrt{\text{var}(mel_7)} \qquad (1)$$

where $\overline{mel}_7$ représents the mean of the 7th Mel filter (mel_7).

As explained previously, many experimental attempts were made, but we kept only the parameter having a strong impact on the speakers number.

Results and interpretation

We recall that the principal objective, expected by this research work, is the estimation of the number of speakers speaking simultaneously during an interview or multi-conference.

For that reason, we have tested the PENS parameter on seven different subsets of the ORATOR database (Ouast 2002), containing seven speech files of 8 seconds each. The experiments are divided into 2 series:

First evaluation: estimation of the number of speakers

This evaluation experiment consists in trying to estimate the number of speakers speaking simultaneously in a sequence of a speech file. Results of estimation are presented in figure 1.

We notice on figure 1 that we can easily know if the speech file contains the speech of only one speaker or a speech mixture of two speakers or more.

So, we can easily deduce if only one speaker is speaking or if it is a speech overlap of different speakers; this estimation is then accurate with a precision of 100% since the separation range between the files of a single speaker and the other cases is considerable.

For the files containing three mixed speakers, the estimation error is 14.3%. This error increases if the number of speakers exceeds 4 speakers (figure 1).

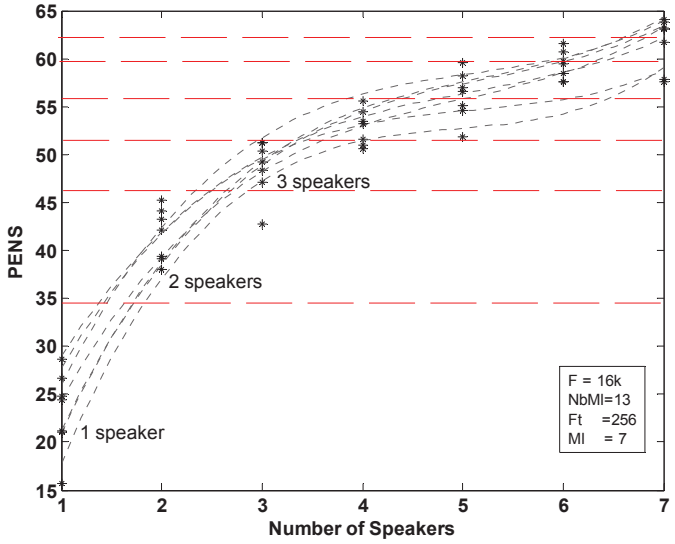

Figure 1: PENS versus the number of speakers.

Second evaluation: test of discrimination

This test consists in making a discrimination between two speech signals according to the speakers number. Results of discrimination, presented in table 2, show that the discrimination between an audio document containing the speech of a unique speaker and another one containing the speech of two speakers can be made without any ambiguity and without any error. We get the same result if the difference between the speakers number is more than one (e.g. between 2 and 4 speakers or between 3 and 5 speakers, etc…).

Table 2: Discrimination according to the number of speakers in %.

Discrimination	Good discrimination	Error of discrimination
1 and 2 speakers	100	0
1 and several speakers	100	0
2 and 3 speakers	92.9	7.1
2 and 4 speakers	100	0
3 and 4 speakers	85.7	14.3
3 and 5 speakers	100	0
4 and 5 speakers	78.6	21.4
4 and 6 speakers	100	0

Conclusion and interpretation

The objective of this work is to find a confidence parameter allowing us, in one hand to distinguish a mono-speaker speech segment from a multi-speaker speech segment, and in the other hand, to estimate the number of speakers sharing a discussion simultaneously with the lowest error possible. Experimental results show that the use of the new confidence parameter PENS gets an interesting precision. The discrimination between single and multi-speaker segments has an error of 0%, and the estimation of the number of speakers, talking in same time, has an error of 0% for a unique speaker, an error of 0% for two speakers and an error of 14,3% for three speakers. Over four speakers, the estimation becomes less accurate.

Finally, we consider that this research domain remains little explored even though the problems of speech overlap, encountered in practice, are very restrictive in speech recognition or audio indexing.

References

Arai, T. 2003. Estimating Number of Speakers by the Modulation Characteristics of Speech. ICASSP, 197-200.

Asano, F., Yamamoto, K., Ogata, J., Yamada, M. and Nakamura, M. 2007. Detection and separation of speech events in meeting recordings using a microphone array. EURASIP Journal on Audio, Speech, and Music. Volume 2007, ID 27616.

Overlapping speech. http://changingminds.org/
techniques/conversation/interrupting/overlap_speech.htm

Lee, H.S. and A.C. TSOI. 1995. Application of multi-layer perceptron in estimating speech / noise characteristics for speech recognition in noisy environment. Speech Com. 17, 59-76.

Quast H. 2002. Automatic recognition of nonverbal speech. An Approach to model the perception of para- and extraling. Vocal Commun. with Neural Net. Mach. Per. Inst. for Neural Comput. UC San Diego, June 28 2002.