Journal of Information Hiding and Multimedia Signal Processing ©2010 ISSN 2073-4212
Ubiquitous International Volume **1**, Number **2**, April **2010**

Proposal of a New Confidence Parameter Estimating the Number of Speakers -An experimental investigation-

Halim Sayoud and Siham Ouamour

Institute of Electronics and Computer Engineering
USTHB University
BP 32 Al-Alia, Bab-Ezzouar, Alger, Algeria
halim.sayoud@gmail.com; siham.ouamour@gmail.com

Received January 2010; revised March 2010

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 systems of automatic speaker recognition. In practice, these systems present a strong degradation in such situations. 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, where each set contains seven multi-speaker files. Results show that the 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.*
Keywords: speech processing, speaker recognition, speech overlap, estimation of the number of speakers.

1. **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 [1]. Thus, the speech signal will contain some zones of speech overlap: F. Asano [2]. Such cases often arise with female speakers: according to [3], women have a multi-task behavior which permits them to speak and understand in such conditions, although that case may also arise with male speakers, often during hot debates between adversaries presenting opposite ideas, such as political debates for example.

Moreover, those speech overlaps may characterize specifically one language more than another: According to [3] for instance, in certain regions of Italy 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 [4], 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. The paper is structured as follows: we will present the used speech database in section two, describe the new algorithm in section three and show the experimental results in section four. We will conclude at the end of the paper by giving some interpretations and discussions on the results.

2. **Database.** The speech database is a subset of ORATOR database: Holger Quast [5], developped by Holger Quast in 2002 ("Machine Perception Laboratory -Institute for Neural Computation of San Diego" and "Drittes Physikalisches Institut - Georg August Universitt Gttingen"). It consists in German talks uttered by real actors expressing various emotional states: anger, joy, etc. It was recorded in 1998 (November/ December) in Gttingen and Northeim, Germany The list of the speakers used in ORATOR is listed here below [5]. Note that:
- the letter A: denotes an Actor, - the letter M: a Male speaker, - the letter F: a Female speaker, - and the letter N: denotes a Non-actor speakler.
List of the speakers of ORATOR database:

- AM1, age =27, Bavarian accent;
- AM2, age =38, slight Swiss accent;
- AM3, age =40;
- AM4, age =27;
- AM5, age =36;
- AM6, age =66;
- AM7, age =56;
- NM1, age =24, student and banker;
- NM2, age =41, worker;
- NM3, age =39, worker;
- NM4, age =38;
- NM5, age =31, group manager;
- NM6, age = over 30, worker;
- NM7, age =57, worker;
- NM8, age = 55, salesman;
- NM9, age =25, student;
- NM10, age =72, professor of physics;
- NM11, age =57, teacher;
- NM12, age =24, student, banker;
- AF1, age =29;
- AF2, age =33;
- AF3, age =34;
- AF4;
- AF5, age =over 30;
- AF6, age =28;
- NF1, age =21, student;
- NF2, age =55, home/ loans specialist.

For this purpose, 145 spoken versions of the following German sentences were recorded [5]:

- *In der Vergangenheit ist schon einiges an guter Vorarbeit geleistet worden.*
- *Die Ziele, die wir jetzt verfolgen, sind die gleichen und mssen auch auf die gleiche Weise*

behandelt warden.
- Unsere Aufgabe ist nun, noch einmal die Zeiteinteilung durchzusehen.
- Sie berprfen dann das Weitere.
- Bitte notieren Sie die Punkte, die Sie heraussuchen, und tragen Sie uns diese vor!
- Wir erledigen alles Andere.
- Glauben Sie, dass Sie das schaffen?
- Gut!

In our experiments, we did not use the entire corpus but have limited our experimental speech database to some female and male speakers present in ORATOR. So, the speech signals uttered by the different speakers of the database, are mixed to get a composite signal representing the speech overlap (see figure 1). Thus, we have chosen some utterances belonging to the selected speakers and built seven subsets for the experimental test, as described below. List of the experimental subsets we have built:

- Mixed Differ Txt: the speakers pronounce different texts;
- Mixed Differ Txt2: the speakers pronounce different texts;
- Mixed Same Txt: the speakers pronounce the same text;
- Mixed Same Txt2: the speakers pronounce the same text;
- Mixed Differ Txt3 M: male speakers pronounce different texts;
- Mixed Differ Txt4 F: female speakers pronounce different texts;
- Mixed Same Spk: the same speaker pronounce different texts.

Each of the 7 previous subsets contains 7 speech files of 8 seconds (length of one speech file), sampled at 16 kHz. The acoustic processing is done on segments of 2 seconds, with a FFT based spectral analysis on 16 ms and a gaussian MEL filtering with 13 filters. The seven files considered in each sub-set contain respectively a file with 1 speaker, a file with 2 speakers, a file with 3 speakers, and a file with 7 speakers talking simultaneously. Thus, in the overall, we have 49 files of speech to process.

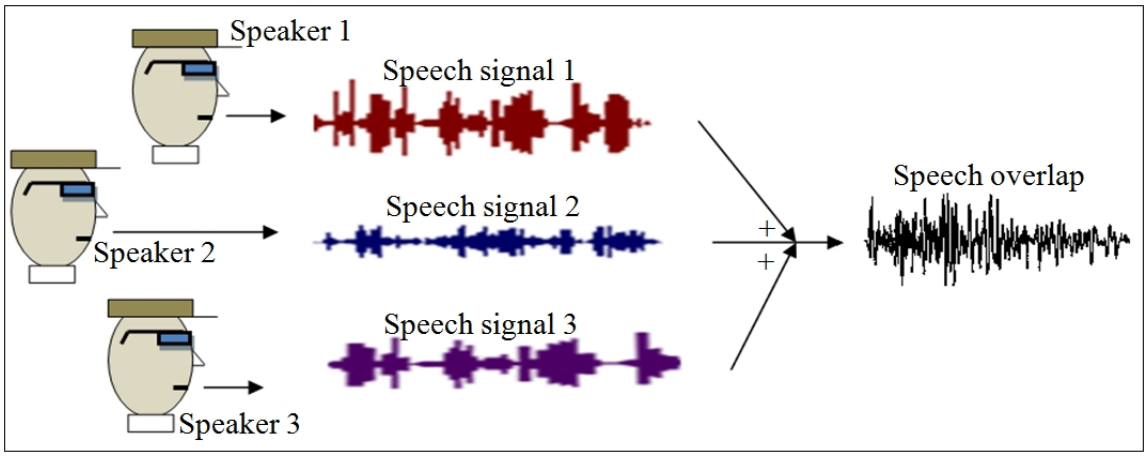

FIGURE 1. The speech signals, uttered by 3 different speakers, are mixed to get a composite signal representing the speech overlap.

3. **Approach Description.** This research work deals with the estimation of the number of speakers in a speech mixture, as described in the works of Takayuki Arai [1]. Our approach is based on the statistical characteristics of the 7th Mel filter, as described in the research works of H. S. Lee and A.C. TSOI [6]. The Mel filter has a Gaussian form with a maximum of amplitude located at the median frequency, which is equal to 1.8375

kHz. It has a right cut-off frequency at 50% of 2.0813 kHz and a left cut-off frequency at 50% of 1.6125 kHz. The spectrum characteristics of that Mel filter are presented in table 1 and figure 2.

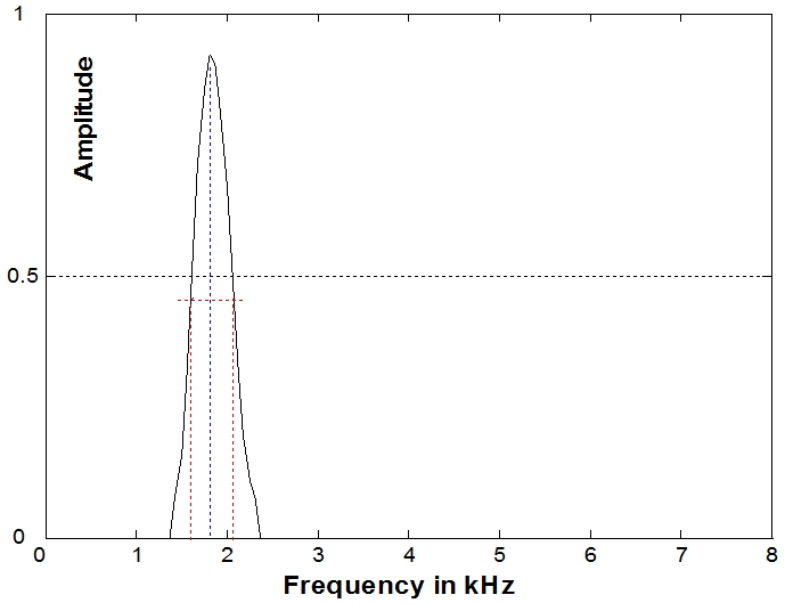

FIGURE 2. Spectral form of the 7th Mel filter (in a melbank of 13 Mels)

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). The PENS parameter is given by:

$$PENS = \overline{mel_7} - \sqrt{var(mel_7)} \tag{1}$$

where $\overline{mel_7}$ represents the mean, and $var(mel_7)$ represents the variance of mel_7.

As explained previously, many experimental attempts were made, but all the parameters were discarded except that having a strong impact on the estimation of the number of speakers.

4. **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 a multi-conference or interview. For that reason, we have tested the new parameter PENS on 7 different subsets that are extracted from the ORATOR database, which was developed by Holger Quast [5]. Our experimental database contains 49 speech files of

8 seconds each (see section 2). Thus, the different results obtained are represented in figures 3 to 5, and tables 2 and 3. The experimental evaluations are divided into 2 series of experiments: a first evaluation estimating the number of speakers and a second evaluation making some tests of speakers' number discrimination: allowing us to distinguish between the different speech files according to the number of speakers present in each file (eg. making a distinction between mono-speaker speech an bi-speaker speech).

4.1. **First evaluation: Estimation of the number of speakers.** This evaluation test consists in trying to estimate the number of speakers speaking simultaneously in a sequence of a speech file. Results of estimation are presented in figures 3, 4 and 5, and in table 2. We notice on figure 3 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 of the speakers' number is 14.3%. This error (Er) is calculated by the ratio between the number of false estimations and the total number of speech files of the assessment:

$$Er = \frac{number\,of\,false\,estimations}{total\,number\,of\,estimations} \tag{2}$$

It increases if the number of speakers present on the file exceeds 4 speakers (see table 2). In this case, we can notice that the error reaches 28.6%. Figure 3 represents the curves giving the PENS average (for the 49 experiments) versus the number of speakers. This curve is approximated by polynomial interpolation and is given by:

$$PENS(x) = 0.37x^3 - 5.62x^2 + 30.37x - 1.26 \tag{3}$$

with x represents the number of speakers.

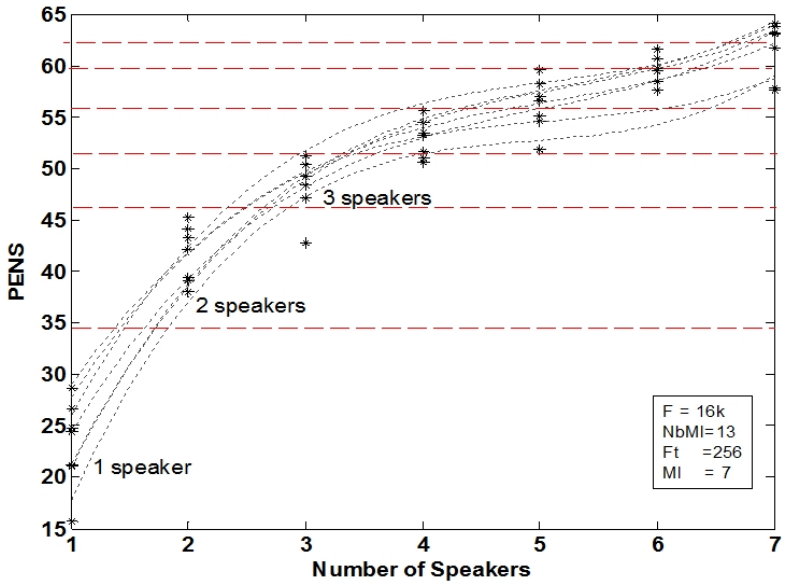

FIGURE 3. PENS versus the number of speakers.

PENS stands for "Parameter Estimating the Number of Speakers".

A mathematical interpolation allows us to estimate the expression of the average polynomial. The expression of this polynomial is given by the equation 3.

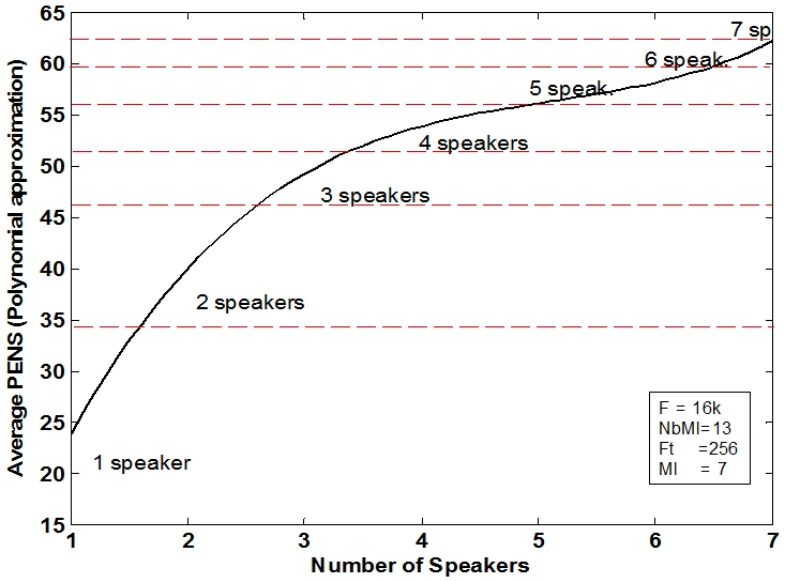

FIGURE 4. Polynomial expression of PENS versus the number of speakers.

TABLE 2. Good estimation of the speaker's number in %

Number of speakers speaking simultaneously	Good estimation of the speakers number in %	Estimation error in %
1	100	0
2	100	0
3	85.7	14.3
4	71.4	28.6
5	57.1	42.9
6	42.9	57.1
7	57.1	42.9

In table 2 and figure 5, we can observe that the estimation of one speaker or two speakers is made without any error. Over four speakers speaking simultaneously, the estimation becomes difficult (error superior to 28.6%). Figure 5 shows that the score of good estimation decreases linearly when the number of real speakers increases, except for the last case (7 speakers) where the score increases paradoxically. However, this phenomenon is not significant.

4.2. **Second Evaluation: Test of Discrimination.** This experiment consists in trying to do a test of discrimination between two speech signals according to the number of speakers. In fact, we made some experiments of discrimination between signal files containing "n" speakers (who are speaking simultaneously) and "m" speakers (who are speaking simultaneously). The discrimination method consists then in deciding whether n and m are equal or not: for example, checking whether an audio signal is mono-speaker or bi-speaker. This process is done automatically thanks to the PENS criterion. The results of discrimination, presented in table 3, show that the discrimination between an audio document containing the speech of a unique speaker and another one containing the speech of two speakers is done without any ambiguity and without any error. We get

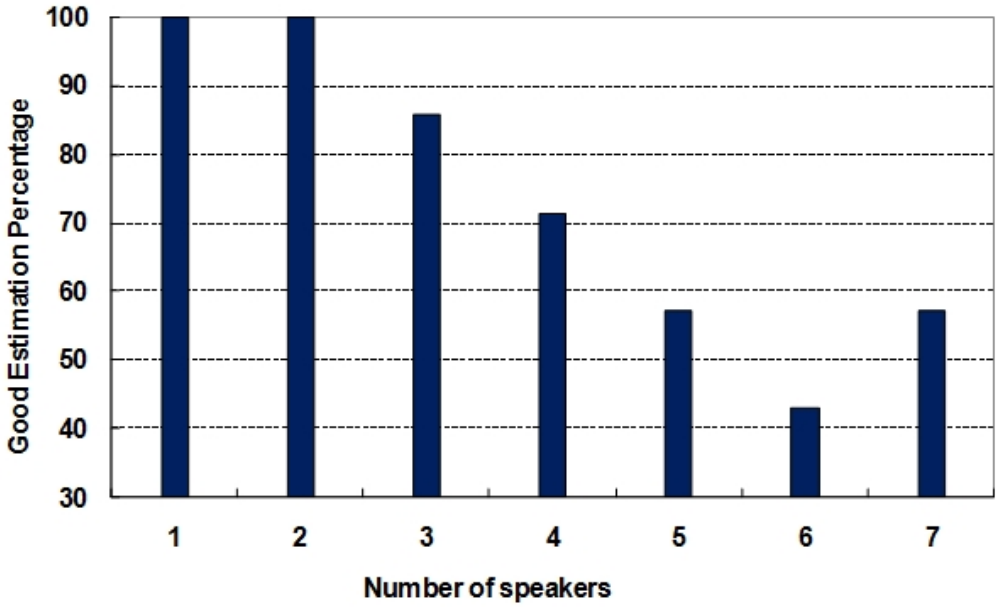

FIGURE 5. Good estimation of the speakers' number in % versus the real number of speakers.

the same observation for the discrimination between n speakers and m speakers if the difference between the numbers n and m is greater than one (e.g. between 2 and 4 speakers or between 3 and 5 speakers, etc.)we can discriminate them without any error. In other words, this means that our system can separate speech signals presenting a difference in the number of speakers that is equal or greater than 2 speakers. Thus for example, we manage to discriminate a signal containing the speech mixture of 3 speakers and a signal of 5 speakers, or between a speech signal of 2 speakers and a speech signal of 4 speakers (table 3). And particularly for the application of speech overlap detection, the system gives good performances (100% of good detection), in the case of distinguishing between the signals containing only one speaker and those containing more than one speaker (table 3).

In the overall, according to this table, we notice that the discrimination between 1 and 2 speakers is made without any error. In the other cases, the discrimination is generally less accurate.

TABLE 3. Discrimination according to speaker's number in %

Discrimination between:	Good discrimination rate %	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

5. **Conclusion and discussion.** One major problem encountered during the speech analysis for applications related to speech recognition, speaker recognition or automatic speech indexing [7], is the difficult signal processing within the areas presenting a mixed speech, namely: speech segments containing a speech mixture of two or several speakers, speaking in the same time (as commented in the works of A. Quinlan and F. Asano [8]). The objective of this work, then, is to find a confidence parameter allowing us, on one hand to distinguish a mono speaker speech segment from a multi-speaker speech segment, and on the other hand, to estimate the number of speakers sharing a discussion simultaneously with the lowest error. The different experimental results show that the new confidence parameter PENS, based on the statistical characteristics of the 7th MEL coefficient, permits us to get an interesting precision of speakers number estimation. The discrimination between single and multi-speaker segments has an error of 0%, and the estimation of the number of speakers talking simultaneously 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 (error of 28.6%). This fact shows the efficiency of our confidence parameter PENS in the separation of mono-speech areas from speech overlap areas. Moreover, this parameter permits us to make also a discrimination between some categories of speech signals with regards to the number of speakers. Concerning our perspectives, we wish to encourage more investigations in this field, since this research domain appears to be not very explored in the different topics of speech processing, even though the problems of speech overlap that are encountered in practice are very restrictive for the systems of speech and speaker recognition or audio indexing [9]. Finally, by this modest research work, we hope to bring some statistical techniques that can be used in speech or speaker recognition, in order to enhance the recognition task.

Acknowledgments. We would like to thank all the persons who contributed, directly or indirectly, to our research achievements. A particular part of acknowledgements to Dr Holger Quast from the Machine Perception Laboratory (Institute for Neural Computation of San Diego) and Georg August Universitt Gttingen University. We wish also to thank the reviewers and the editors who accepted to read this paper.

REFERENCES

[1] T. Arai, Estimating number of speakers by the modulation characteristics of speech, *Proc. of ICASSP*, pp. 197-200, 2003.

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

[3] Dave, *Overlapping Speech*, http://changingminds.org/techniques/conversation/interrupting/overlap_speech.htm

[4] S. Ouamour, *Indexation Automatique des Documents Audio en vue d'une Classification par Locuteurs - Application l'Archivage des missions TV et Radio*, PhD thesis, ENS. Polytechnique. October 1st 2009.

[5] H. Quast, *Automatic Recognition of Nonverbal Speech. An Approach to Model the perception of Para- and Extralinguistic Vocal Communication with Neural Networks*, Thesis. University of Gottingen, 2001.

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

[7] S. Ouamour, M. Guerti and H. Sayoud, Speaker based segmentation on broadcast news - on the use of ISI technique, *Proc. of ISCA Tutorial and Research Workshop on Experimental Linguistics*, pp 193-196, Athens, Greece, August 2006.

[8] A. Quinlan and F. Asano, Detection of overlapping speech in meeting recordings using the modified exponential fitting test, *Proc. of the 15th European Signal Processing Conference (EUSIPCO 2007)*, Poznan, Poland, pp. 3-7, September 2007,

[9] S. Steininger, F. Schiel, and K. Louka, Gestures during overlapping speech in multimodal human-machine dialogues, *Proc. of International Workshop on Information Present-ation and Natural Multimodal Dialogue*, Verona, Italy, 2001.

[10] S. Ouamour, H. Sayoud and M. Guerti, PENS: A confidence parameter estimating the number of speakers, *Proc. of ISCA Tutorial and Research Workshop on Experimental Linguistics*, pp. 161-164, Athens, Greece, August 2008.