

53. IWK

Internationales Wissenschaftliches Kolloquium
International Scientific Colloquium



Faculty of
Mechanical Engineering



PROSPECTS IN MECHANICAL ENGINEERING

8 - 12 September 2008

www.tu-ilmenau.de

th
TECHNISCHE UNIVERSITÄT
ILMENAU

Home / Index:

<http://www.db-thueringen.de/servlets/DocumentServlet?id=17534>

Published by Impressum

Publisher
Herausgeber Der Rektor der Technischen Universität Ilmenau
Univ.-Prof. Dr. rer. nat. habil. Dr. h. c. Prof. h. c. Peter Scharff

Editor
Redaktion Referat Marketing und Studentische Angelegenheiten
Andrea Schneider

Fakultät für Maschinenbau
Univ.-Prof. Dr.-Ing. habil. Peter Kurz,
Univ.-Prof. Dr.-Ing. habil. Rainer Grünwald,
Univ.-Prof. Dr.-Ing. habil. Prof. h. c. Dr. h. c. mult. Gerd Jäger,
Dr.-Ing Beate Schlütter,
Dipl.-Ing. Silke Stauche

Editorial Deadline
Redaktionsschluss 17. August 2008

Publishing House
Verlag Verlag ISLE, Betriebsstätte des ISLE e.V.
Werner-von-Siemens-Str. 16, 98693 Ilmenau

CD-ROM-Version:

Implementation
Realisierung Technische Universität Ilmenau
Christian Weigel, Helge Drumm

Production
Herstellung CDA Datenträger Albrechts GmbH, 98529 Suhl/Albrechts

ISBN: 978-3-938843-40-6 (CD-ROM-Version)

Online-Version:

Implementation
Realisierung Universitätsbibliothek Ilmenau
[ilmedia](#)
Postfach 10 05 65
98684 Ilmenau

© Technische Universität Ilmenau (Thür.) 2008

The content of the CD-ROM and online-documents are copyright protected by law.
Der Inhalt der CD-ROM und die Online-Dokumente sind urheberrechtlich geschützt.

Home / Index:

<http://www.db-thueringen.de/servlets/DocumentServlet?id=17534>

S. Pleshkova-Bekiarska / D. Damyanov

Spectral transform technique for speech identification and overlap detection

ABSTRACT

Speech identification is a very interesting topic nowadays with important applications in the security systems, where reliable and safe speech identification should be performed. Unfortunately, a lot of distortion factors make this process extremely difficult to perform. It is not always possible that only the voice of the person that is to be identified is recorded. It is very often the case, that multiple voices are recorded. Then, the recognition system should know, whether it is possible to identify the voice from one speaker, or the energies of all of the speakers are equal, and no recognitions could be performed. In this case, our spectral transform algorithm comes into play. When the recorded mixture of voices is spectrally transformed, it is possible that the target speaker can be identified. Our spectral technique includes LPC coefficient extraction, then a model of the target speaker is developed. This model contains all of the spectral features, which we expect from the source talker. Finally, an optimization of the conversion LPC spectrum to LPC coefficients is performed, so that the model of the spectrum of the voice from the target talker minimally differs from the optimized one. Going this way it is possible to identify the target talker from a voice record, containing voices of more than one speaker.

I. INTRODUCTION

Speech overlap is the simultaneous occurrence of speech from more than one speakers. It has some very bad effects in the work of speech recognition systems. Speech overlap detection is one of the main areas in speech and speaker indexing. In speaker indexing, speech signal is partitioned into segments where each segment is uttered by only one speaker. So, parts of speech that include two or more

speakers simultaneously should be determined before any following processes. Speaker overlap detection is also useful in some other speech processing applications including speech and speaker recognition. In this paper the methods, a new method for speech identification and overlap detection is introduced. There are some traditional methods for performing the upper such as Spectral Auto-Correlation Peak Valley Ratio (SAPVR) and the K-nearest method (KNN) [1] [2] [3] [8]. They have some advantages in the area of blind source separation, i.e. when we have a mixture of two or more apriori unknown speech signals. There are also some new methods for overlap detection and speaker indexing, such as the ones, which use high-order statistics. This paper concerns about the case, when it is possible to record samples of the speech of each speaker, which voice is then mixed with the others. Is it clear to see, that when some information about the speakers is available, the method, proposed in this paper, performs much better than the others.

II. GENERAL DESCRIPTION OF THE PROPOSED SPECTRAL TRANSFORM TECHNIQUE (STF)

The description of the algorithm goes as following. First, some speech of each speaker is recorded. Every one of the speakers therefore says one and the same sentence of speech which is the recorded. Then the voices of the speakers are recorded as they speak simultaneously, each of them saying one and the same sentence, which is the hardest possible case. The speakers are situated at 1 meter distance to each other. This is done for recognition sake. The algorithm for overlap detection is implemented in a system, that is embedded in a small robot, together with a microphone. The position of the robot is then fixed, so that he is forced to record and process a voice from only one speaker. This means that the robot is designed to understand the speech of only one speaker at a time, while the other are talking and so producing overlapping. So the robot records then the speech from each speaker, as the others are talking. After that each speech signal is processed as follows. First, its spectral features are extracted, i.e. its Fourier spectrum is evaluated [4,7]:

$$S(w) = \int_{-\infty}^{+\infty} S(t)e^{-j\omega t} dt \quad (1),$$

in the continuous way, where w is the frequency, and t denotes the time, and

$$S(n) = \frac{1}{N} \sum_{k=0}^{N-1} S(k) e^{j2\pi nk/N} \quad (2),$$

where n is the discrete frequency, k is the discrete time, N is the number of samples.

The algorithm uses windowing of the signal, before the spectral features extraction, for some very well known reasons. After the spectrum is available, for the spectrum of each windowed part the LPC coefficients are evaluated. The evaluation of the LPC coefficients is done by the well known way. The coefficients of the predicting filter are evaluated [5]:

$$A(z) = \sum_{i=0}^{NP} a_i Z^{-i} \quad (3),$$

Where a_i is

$$\begin{bmatrix} r_0 & r_1 & \dots & r_{NP-1} \\ r_1 & r_2 & \dots & r_{NP-2} \\ \cdot & & & \\ \cdot & & & \\ r_{NP-1} & r_{NP-2} & \dots & r_0 \end{bmatrix} * \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_{NP} \end{bmatrix} = \begin{bmatrix} -r_0 \\ -r_1 \\ \cdot \\ \cdot \\ -r_{NP-1} \end{bmatrix} \quad (4),$$

Where r_i are the coefficients of the autocorrelation matrix and are evaluated as follows:

$$r_i = \sum_{n=i}^{NF-1} \tilde{S}(n) * \tilde{S}(n-i) \quad (5)$$

Where $\tilde{S}(n)$ - are the values of the speech signal in one window, with $h(n)$ being the window function :

$$\tilde{S}(n) = S(n) * h(n) \quad (6), \quad (4)$$

$$h(n) = 0,54 - 0,46 * \cos\left(\frac{2\pi n}{NF-1}\right) \quad (7).$$

Then again, the Fourier spectrum is evaluated, but not for the speech signal, but for the LPC coefficients. This means that the LPC coefficients are viewed as a sampled signal and its spectrum is evaluated. Then, a mean arithmetic value of all of the LPC-spectra of all speakers is evaluated. The for each recorded from the robot overlapped speech, the Fourier spectrum is found. Then the spectrum is processed with the average LPC coefficients, and then the new averaged speech signal is evaluated [5]:

$$LPC_{arithmetic\ mean} = \frac{1}{M} \sum_{i=1}^M LPC_i \quad (8)$$

With M being the number of speakers.

This procedure is repeated for each position of the robot, i.e. for the recorded signal of each speaker, while the others are talking and thus causing overlapping. Finally, the speech signal of the speaker, speaking alone, is compared to the processed by the algorithm overlapped signal for the particular speaker. Then the mean square is evaluated. This mean square error is then compared to the existing methods [6,8]:

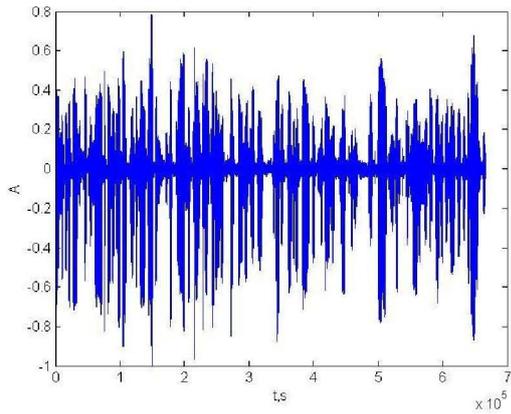
$$MSE = \frac{1}{N} \sum_0^{N-1} |S(n) - \hat{S}(n)|^2 \quad (9)$$

Where $\hat{S}(n)$ is the processed speech signal.

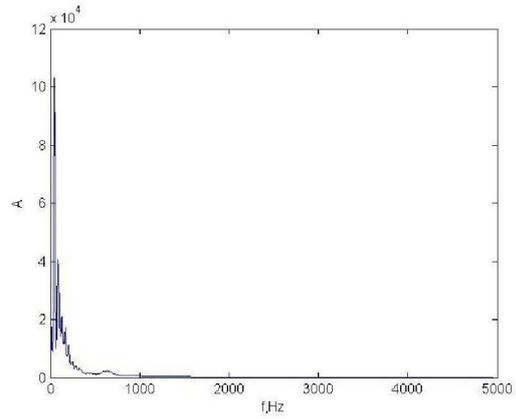
It can be observed, that the error is much smaller with the proposed method, because of the information about the speakers, that we apriori have. The other methods have bigger errors, for they are working with blind source separation.

III. Evaluation of the algorithm and results

One of the signals of the speakers is given at fig.1, and it's Fourier spectrum at fig.2

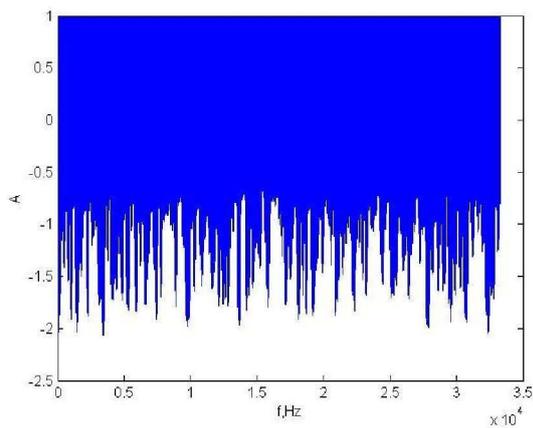


**Fig.1 Speech signal of one speaker
In the discrete time domain**

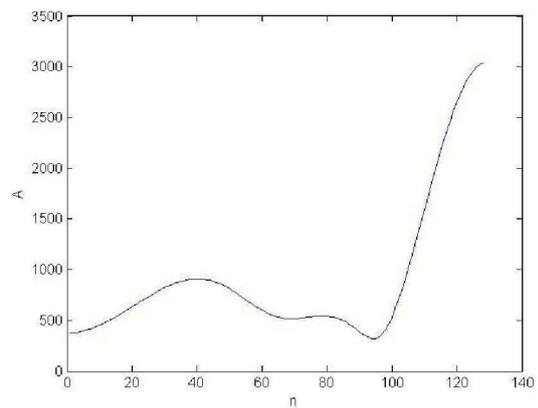


**Fig.2 The Fourier spectrum of
speech signal of the same speaker
as in fig.1**

The LPC coefficients of the upper signal are shown at fig 3, and its spectrum at fig.4



**Fig.3 The LPC coefficients of the
speech of the speaker from fig 1.**



**Fig.4 The Fourier spectrum of
LPC coefficients from fig.3**

The average spectrum of the coefficients is shown at fig. 5

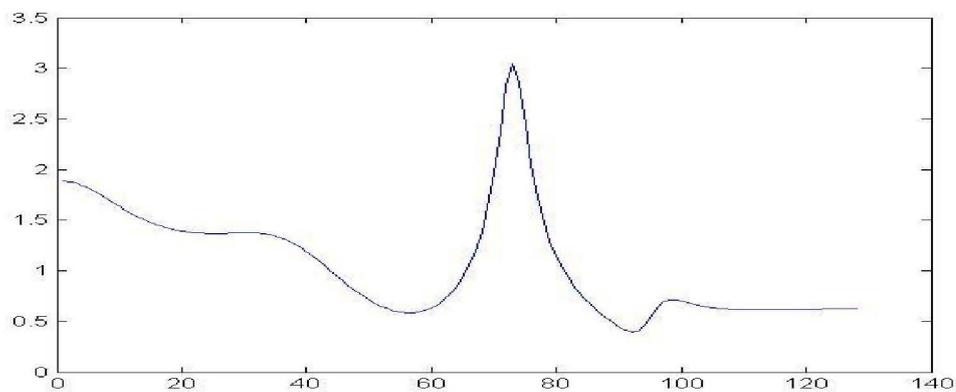


Fig.5 The average of the LPC coefficients of all speakers

The results for the mse, compared to the other methods are given at table 1 [1].

overlapped voices/ method	SAVPR	Third Order Moment Method	STF
female - female	2.4 - 2.4	1.9 - 2.1	1.1 - 1.2
female - male	2.5 - 1.9	2.1 - 1.8	1.2 - 1.3
male-male	2.2 - 1.9	2.0 - 1.8	1.0 - 1.1
single vowels	5.2 - 5.3	5.2 - 5.0	2.5 - 2.6

Table 1

IV. CONCLUSION

The proposed spectral transform technique clearly showed that it performs much better than the other blind source separation algorithms, for it uses some apriori data for the speakers, with voices cause overlap. On direction for future work of the authors is to implement the proposed technique without having the apriori information, and compare the results.

Acknowledgements

This work was supported by National Ministry of Science and Education of Bulgaria under Contract BY-I-302/2007: "Audio-video information and communication system for active surveillance cooperating with a Mobile Security Robot".

REFERENCES

- [1] High-Order Statistics application for speech identification and overlap detection" Pleshkova – Bekiarska Snejana, Damianov Damian, ISRSSP-2007, Sofia, pages 59-63
- [2]Speech Overlap Detection using Spectral Features and its Application in Speech Indexing – M. H. Moattar, M. M. Homayounpour,
- [3] J. M. Mendel "Tutorial on High Order Statistics (Spectra) in signal processing and system theory: theoretical results and some applications *Proceedings of IEEE* 79(3), pp 278-305, March, 1991.
- [4] Digital Signal Processing (Schaum's Outlines, OCR) McGraw-Hill 1999.
- [5] Spiegel, Murray R. "Schaum's outline of theory and problems of statistics". ISBN 0-07-060234-4
- [6] Digital Speech, A.M. Kondoz, 1999.
- [7] R. Arnaudov, R. Miletiev - Inertial sensor data analysis using nonuniform sampling, Proceeding of the International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services – TELSIS'2007, September 26-28, 2007, Nis, Serbia, vol.1, pp.309-312
- [8] A. Bekiarski, Bl. Shishkov, M. Dobрева. High-Order Statistics in Blind Image Restoration. International Symposium on Radio Systems and Space Plasma of International Union of Radio Science (URSI), Sofia, September, 2007, pp. 23-26.

Authors:

Teaching assistant PhD Snejana Pleshkova-Bekiarska
 Teaching assistant Damyan Damyanov
 Technical University of Sofia, Kliment Ohridski Blv 8, Sofia, Darvenitsa, Bulgaria, 1-st block, room 1258.
 1756 Sofia

Phone: 0035929653300

Fax: none

E-mail: snegpl@t-sofia.bg, ellov@abv.bg,