Speech Overlap Detection Algorithms Simulation

Snejana Pleshkova-Bekiarska[1] and Damyan Damyanov[2]

Abstract **– The concern of this paper are different methods of speech overlap detection. Speech overlap is the simultaneous occurrence of speech from more then 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 method for speech overlap detection Spectral Auto-Correlation Peak Valley Ratio (SAPVR) is shown. At the end of this paper, the results from the work of the methods are plotted. They are the precision rate and the detection rate. The average time for processing for 1 second of speech is also taken under consideration.**

Keywords **– Spectral Auto-Correlation Peak-Valley Ratio, K nearest neighbour, speech overlap detection.**

I. INTRODUCTION

The auto-correlation is a standard method of evaluating how correlated is a signal with a copy of itself, delayed on certain interval d. If we have the series x(n) the auto-correlation of this signal is

$$r = \frac{\sum_i [x(i) - m_x] * [x(i-d) - m_x]}{\sqrt{\sum_i [x(i) - m_x]^2} * \sqrt{\sum_i [x(i-d) - m_x]^2}} \quad (1)$$

where m_x is the mean of the series x(i). If the auto-correlation is computed for delays d=0,1,2, …,N-1, then we can write the formula of the auto-correlation with a length twice the length of the signal:

$$r(d) = \frac{\sum_i [x(i) - m_x] * [x(i-d) - m_x]}{\sqrt{\sum_i [x(i) - m_x]^2} * \sqrt{\sum_i [x(i-d) - m_x]^2}} \quad (2)$$

The method of Spectral Auto-Correlation Peak-to-Value Ratio (SAPVR)
uses spectral auto-correlation function to determine whether a speech frame is usable or not [1]. A speech segment is "usable" if it contains enough information to identify the target speaker.
The power spectrum of voiced speech can be predicted because of its harmonic structure. If certain input signals are given , like in fig 1,2 3 and consider a frame of speech that is voiced. The frequency spectrum X(k) of such a frame will contain harmonically related pulses. This operation will always result in pulses of decreasing height with increasing lag. If the original magnitude spectrum X(k) contained harmonics at integral multiples of the digital frequency 'p', then the major contribution to the first peak in the spectral autocorrelation, after lag zero, is due to the product of adjacent harmonics, which occurs at lag 'p'. This is shown in figures 4, 5 and 6.That is, the magnitude of the first spectral peak after lag zero for a voiced frame can be approximated as

$$R(p) = X(p)X(2p) + X(2p)X(3p) + \ldots \quad (3)$$

Other terms will contain less energy, and will not contribute significantly to this peak. Note that this parameter contains all the information about significant
harmonics. The next peak occurs at lag '2p' and its amplitude can be approximated as

$$R(2p) = X(p)X(3p) + X(2p)X(4p) + \ldots \quad (4)$$

By the inherent property of the autocorrelation function, this peak has lesser amplitude than R(p). If the segment of speech is unvoiced, the spectral autocorrelation will not
contain any prominent peaks other than the one at lag 0. [2].The behavior of spectral autocorrelation under co-channel condition varied, depending on whether 1.) both the target and interfering speech were voiced, 2.) either one of them were unvoiced or 3.) both of them were unvoiced. When both the speech frames were unvoiced, the spectral autocorrelation did not contain any pulses that were harmonically related to each other. If at least one of the speech frames was voiced, the spectral autocorrelation contained harmonically related pulses as expected. If both the speech frames were voiced, the spectral autocorrelation contained either two distinct trains of pulses that were harmonically related if the speakers' pitches were different by approximately 25%, otherwise there was one train of broad pulses One important thing is that the ratio of the first local maximum after the one at lag 0, to the local minima between this maximum and the next local minimum, is significantly lower than that of the single speaker case. This is due to the fact that there are significant autocorrelation values for lags that are not harmonically related, due to co-channel conditions. This motivates one to define a spectral autocorrelation ratio, which reflects the extent of corruption of a target speech by the interfering speech.

[1]Snejana Pleshkova-Bekiarka is with the Faculty of Telecommunications,Technical University - Sofia, 8 Kliment Ohridski St.Darvenitsa, 1756, Sofia, Bulgaria, E-mail: snegpl@tu-sofia.bg

[2] Damyan Damyanov is with the Faculty of Telecommunications, echnical University - Sofia, 8 Kliment Ohridski St. Darvenitsa,1756 , Sofia, Bulgaria, E-mail: ellov@abv.bg

The Spectral Autocorrelation Ratio (SAR) parameter is defined as follows:

$$SAR = 20\log_{10}\{R(p_1)/R(q_1)\} \qquad (5)$$

where, R(p1) is the local maximum of spectral autocorrelation other than the one at lag 0 (occurring at lag p1) and R(q1) is the next local maximum that is not harmonically related to the first peak, or the local minimum between p1 and 2p1.
The SAR has to be properly interpreted. If speech of one of the speakers is silent or is unvoiced, a peak that is not harmonically related to the peak due to voicing state
of one talker will be substantially lower in amplitude.This is shown in figures 7,8 and 9. This means the SAR will be very high, from which we would conclude that the frame of speech is usable. If, however,
the speech of target and interferer were of comparable magnitude, the SAR ratio would approach zero, which would identify that particular frame as unusable[3]. What if
there is a spurious peak of comparable magnitude along with the harmonically related pulses? The SAR will again be low, but the physical interpretation is that, a pure tone is
mixed with the speech signal, and if it is of comparable magnitude, that speech frame is definitely unusable.

II. Simulation of Algorithms

The algorithm for evaluating of Spectral Auto-Correlation Peak-to-Value Ratio is as follows:

1. Open and load the wave file in the memory.
2. Create a vector, containing the values of the speech signal.
3. Get the vector length.
4. Create a Hamming window of N points.
5. Evaluate how many windows pass in the vector.
6. For every windowed part of the signal
   - Evaluate the Fourier spectrum
   - Evaluate the spectral autocorrelation

$$r(d) = \frac{\sum_i [x(i)-m_x]*[x(i-d)-m_x]}{\sqrt{\sum_i [x(i)-m_x]^2} * \sqrt{\sum_i [x(i-d)-m_x]^2}} \qquad (6)$$

x(n) , n = 0,1,2, …,N-1

d=0,1,2, …,N-1

   - Estimate the firs peak and second lag after the first lag.
   - Evaluate the SAPVR - $SAR = 20\log_{10}\{R(p_1)/R(q_1)\}$ (7)
   - Set a threshold 6.23 dB
   - If the SAPVR is above the threshold – the frame is usable
   - If not – the frame is unusable – i.e. there is speech overlap and the speaker cannot be identified.

III. Quality Estimation and Comparison

For quality estimation porpoises:

1. Get an amount of data, for which all of the frames are known (usable and not usable).
2. Use the SAVPR algorithm.
3. With the results from the SAVPR algorithm, evaluate the next formulas

$$DR = \frac{length\ of\ truly\ recognized\ non-usable\ segments}{total\ length\ of\ non-usable\ segments} \qquad (8)$$

$$FAR = \frac{length\ of\ truly\ recognized\ usable\ segments}{total\ length\ of\ usable\ segments} \qquad (9)$$

$$PRC - (1 - FAR) * 100 \qquad (10)$$

DR - Detection Rate
FAR - False Alarm Rate
PRC - Precision of the recognition approach

TABLE I
DETECTION RATE

Speech of a man	1
Speech of a woman	1
Speech overlap	1

TABLE II
PRECISION OF THE SAVPR ALGORITHM

Speech of a man	0.16667
Speech of a woman	0.71429
Speech overlap	0

In the following figures, simple signals are shown for visualization purposes. The authors have made an extensive search with many male and female voices.

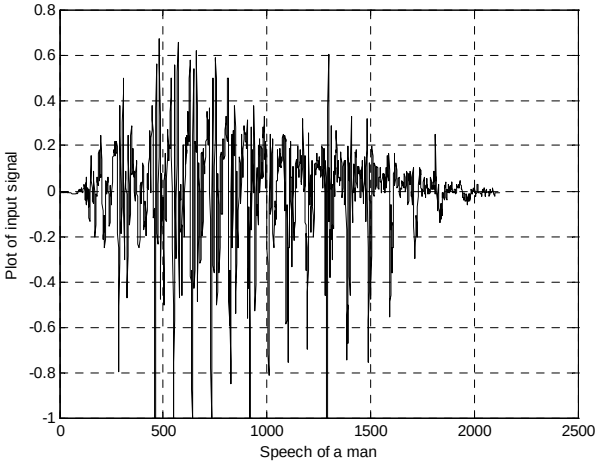

Fig.1. Input signal – speech of a man

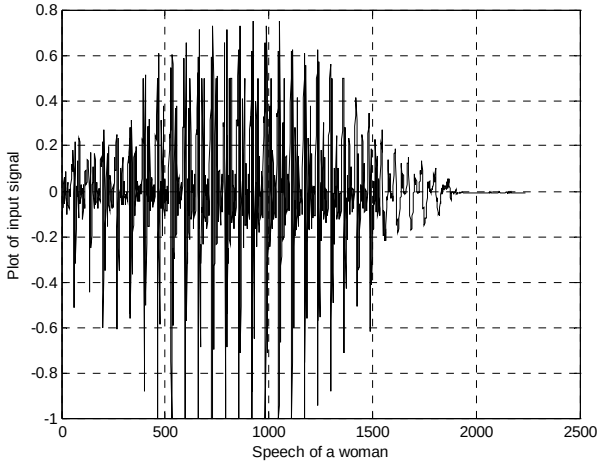

Fig.2. Input signal – speech of a woman

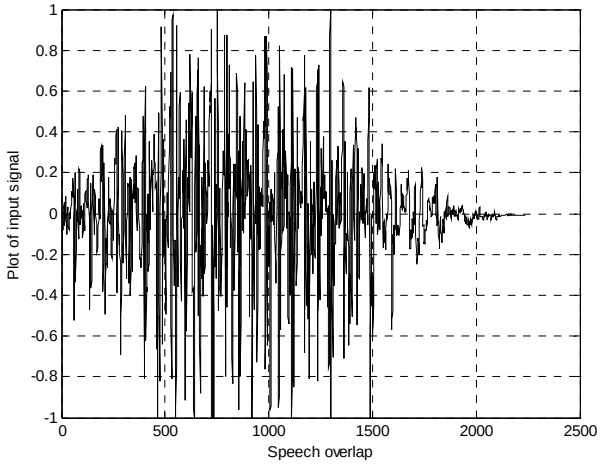

Fig.3. Input signal – speech overlap

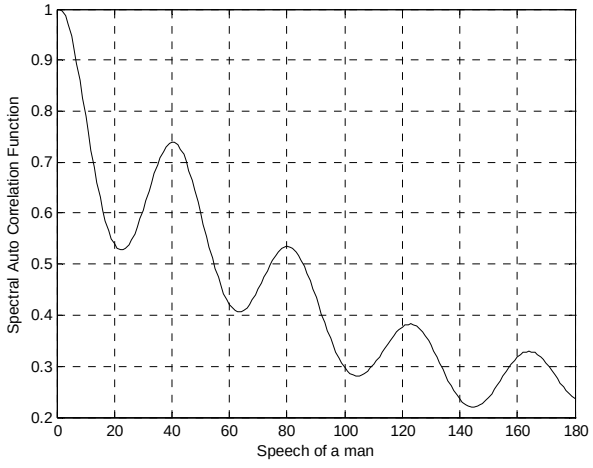

Fig.4. Spectral Auto-correlation – speech of a man

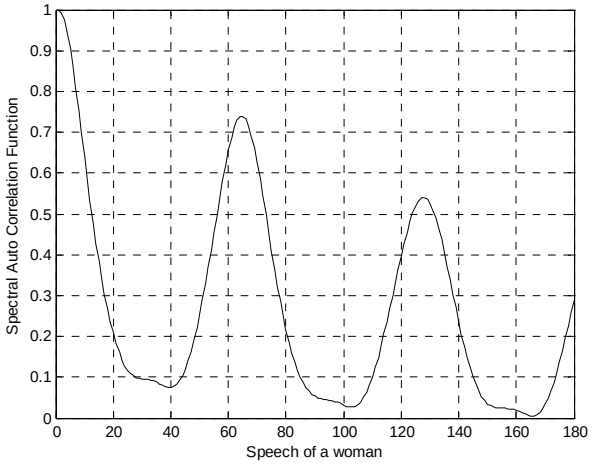

Fig.5. Spectral Auto-correlation – speech of a woman

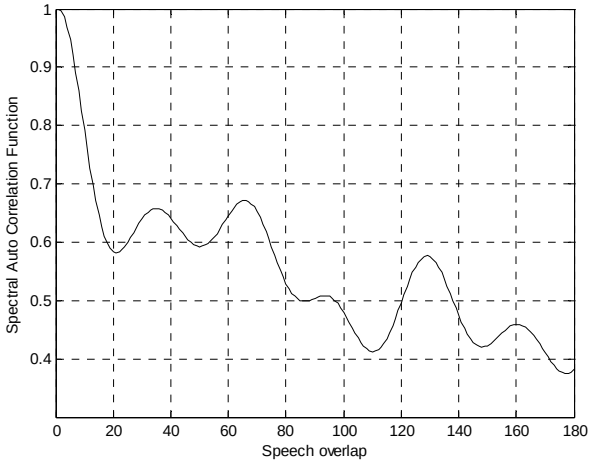

Fig.6. Spectral Auto-correlation – speech overlap

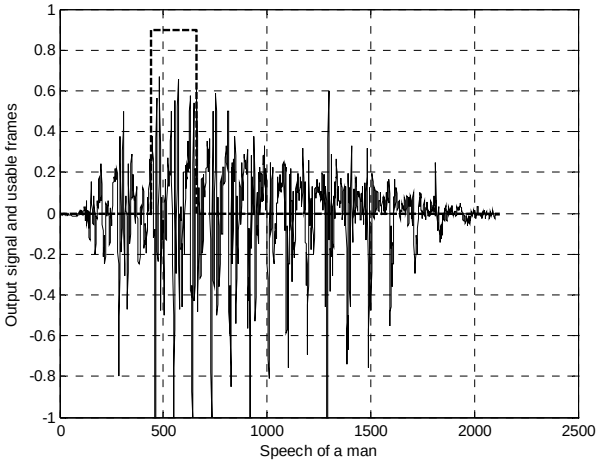

Fig.7. Output signal and usable frames – speech of a man

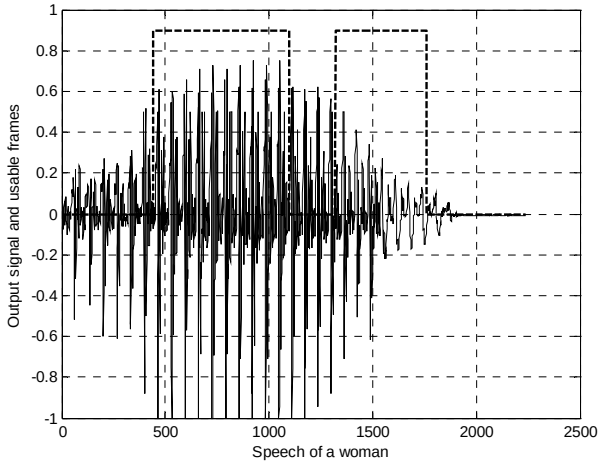

Fig.8. Output signal and usable frames – speech of a woman

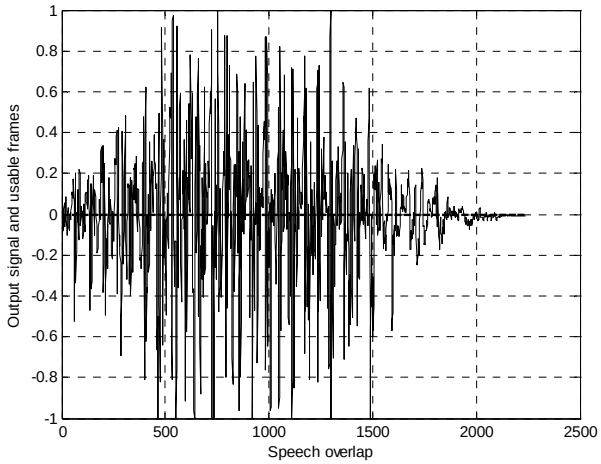

Fig.9. Output signal and usable frames – speech overlap

IV. CONCLUSION

In their future work, the authors have the goal to simulate the other methods, make an appropriate algorithms for them, and show compare the results of the different methods.

REFERENCES

[1] Katsuri Rangan Krishnamachari, Robert E. Yantoro, Damiel S. Benincasa, Stanlet J. Wenndt “Spectral Autocorrelation ratio as a usability measure of speech segments under co-channel comditions ”, IEEE International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2000.

[2] Robert E. Yantoro, "A study of the spectral autocorrelation peak valley ratio (SAVRP) as a method for Identification of usable speech and detection of co-channel speech". AFOSR Rome Labs Summer 2000 Report.

[3] M. H. Moattar, M.M. Homayounpour “Speech Overlap Detection using Features and its Applications in Speech Indexing” Information and Communication Technologies, 2006. ICTTA '06. 2nd.