A NON-UNIFORM REAL-TIME SPEECH TIME-SCALE STRETCHING METHOD

Adam Kupryjanow and Andrzej Czyżewski
Multimedia Systems Department, Gdańsk University of Technology, Narutowaicz 11/12, Gdańsk, Poland

Keywords: Time-scale modification, Voice detection, Vowels detection, Rate of speech estimation.

Abstract: An algorithm for non-uniform real-time speech stretching is presented. It provides a combination of typical SOLA algorithm (Synchronous Overlap and Add) with the vowels, consonants and silence detectors. Based on the information about the content and the estimated value of the rate of speech (ROS), the algorithm adapts the scaling factor value. The ability of real-time speech stretching and the resultant quality of voice were analysed. Subjective tests were performed in order to compare the quality of the proposed method with the output of the standard SOLA algorithm. Accuracy of the ROS estimation was assessed to prove its robustness.

1 INTRODUCTION

Time-scale modification algorithms have been widely investigated by many researchers over last 25 years. Mainly this issue was considered in terms of maximizing the quality of synthesized speech (Moulines, 1995), reduction of computational complexity or its adaptation for real-time signal processing (Pesce, 2000). In this work the main stress was put on design and evaluation of the algorithm which will be able to stretch the speech signal in a real-time, whilst preserving the general synchronization of the original and modified signal. Synchronization is obtained here by the reduction of redundant information in the input signal i.e. shortening of silence and vowels prolongation intervals, stretching vowels and consonants with a different stretching factors and adjusting stretching factor value according to the actual ROS (Rate of Speech).

The proposed algorithm, named Non-Uniform Real-Time Speech Modification algorithm (NU-RTSM), was designed to improve the perception of speech by people with the hearing resolution deficit. It was shown in Tallal's work that the reduction of the speech speed improves its intelligibility (Tallal, 1996). Authors of this paper had proposed the idea of the real-time speech stretching using mobile devices (e.g. Smartphone). Results of that work were described in the conference paper (Kupryjanow, 2010). Some improvements of that algorithm are proposed, i.e. usage of non-uniform time-scaling, in this paper.

As it was shown by Demol (Demol, 2005), non-uniform time-scaling algorithm can improve the quality of processed signal. The assumption of his work was based on the idea that every unit of speech such as: vowels, consonants, plosives, phones transitions and silence should be time-scaled using different scaling factors. Differences between factors were implicated by the prosody rules. Realization of that algorithm is impossible in real-time conditions, because of the problem with the synchronization of the input and output signal (there is no mechanism for the reduction of redundant signal content). In this paper such a mechanism is proposed and examined.

Owing to the structure of the algorithm, it could be implemented on the mobile phone, but because of the legal limitations the processing of the incoming speech stream may be prohibited. Despite the limitations, the modification of the speech could be implemented on the telephone service provider severs or locally on the mobile device working in off-line mode.

2 ALGORITHM DESCRIPTION

In Fig. 1, a block diagram of the NU-RTSM algorithm is presented. The algorithm provides a

Kupryjanow A. and Czyzewski A..
A NON-UNIFORM REAL-TIME SPEECH TIME-SCALE STRETCHING METHOD.
DOI: 10.5220/0003456300270033
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications* (SIGMAP-2011), pages 27-33
ISBN: 978-989-8425-72-0

combination of voice activity detection, vowels detection, rate of speech estimation and time-scale modification algorithms. Signal processing is performed in time frames in the following order:

1. Voice activity detector examines speech presence,
2. For noisy components frame synchronization procedure is performed; if the output signal is not synchronized with the input then noise sample frames are not sent to the output,,
3. Speech sample frames are tested in order to find vowels,
4. Information about vowels locations is used by the ROS estimator to determine the speech rate,
5. Speech frames are stretched up with different stretching factors.

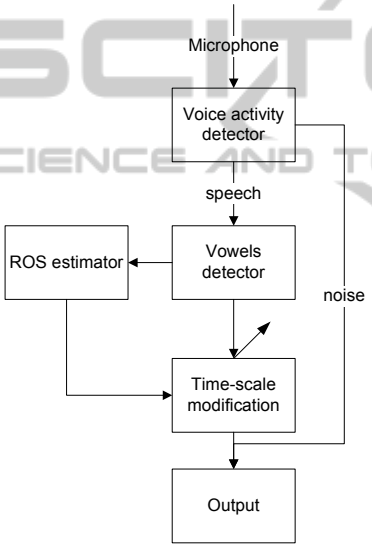

Figure 1: NU-RTSM block diagram.

All algorithms presented in this section were designed and customized to work in a real-time. The input signal for all of them was captured by the headset microphone.

2.1 Voice Activity Detector

Voice activity detection is performed at the beginning of the analysis. The algorithm is designed as a typical VoIP voice detector. Detection of the voice is done in the time frames with a length of 46 ms. For every signal frame spectral energy, defined by Eq. 1, is determined and compared with the energy threshold *Eth*:

$$E = \frac{\sum_{k=1}^{K} A(k)^2}{K} \quad (1)$$

where E represents energy of the frame, $A(k)$ is the k-th spectral line of the input's signal magnitude spectrum and K is the total number of spectrum lines. Energy threshold is obtained at the beginning of the processing, by calculating the mean value of the energy determined for the first 20 frames of the signal. It is assumed that at the beginning of the analysis only noise is recorded by the microphone. Frame is marked as speech if its energy exceeds the *Eth* value.

Threshold value is adjusted to the current noise variations using the two-stage adaptation procedure. First stage is done every time when the frame was marked as noise. For that situation *Eth* is updated using the following formula (Eq. 2):

$$Enth = C((1-p)\cdot Eth + p\cdot E) \quad (2)$$

where *Enth* is the new value of energy threshold, *Eth* is the previous value of energy threshold, E is the energy of the current frame, C is correction factor, and p is the variable which determines how much the new value of the noise energy will influence the value of *Enth.*

The task of the second stage is to fit p value to the actual background noise energy fluctuations. If the variation of the vector that contains last 10 energies, used in the first stage of adaptation, is low, then the energy for the current frame should have low impact on *Eth* adaptation. Therefore, p value is set to 0,2. Otherwise, impact of the current energy should be high, so the p value is set to 0,1.

2.2 Vowels Detector

Vowels detection algorithm is based on the assumption that all vowels amplitude spectra are consistent. To quantify this similarity parameter called *PVD* (peak-valley difference) is used (Moattar, 2010). Initially *PVD* was introduced for the robust voice activity detection. It is defined by the following formula (Eq. 3):

$$PVD(VM, A) = \frac{\sum_{k=0}^{N-1}(A(k)\cdot VM(k))}{\sum_{k=0}^{N-1} VM(k)} - \frac{\sum_{k=0}^{N-1}(A(k)\cdot(1-VM(k)))}{\sum_{k=0}^{N-1}(1-VM(k))} \quad (3)$$

where $PVD(VM, A)$ is the value of peak-valley difference for one frame of the input signal, $A(k)$ is the value of the k-th spectral line of the input's

signal magnitude spectrum and $VM(k)$ is the value of the k-th value in the vowel model vector.

VM is created in the training stage on the basis of the average magnitude spectra calculated for the pre-recorded vowels. The model consists of the binary values, where 1 is placed in the position of the peak in the average magnitude spectrum and 0 for all other positions. When the magnitude spectrum of the input signal is highly correlated with the vowels spectra, *PVD* value is high. Therefore, for the vowels *PVD* takes higher values than for consonants or silence parts.

Vowels detection is executed only for speech frames. Algorithm is based on time frames with the duration of 23 ms. Each signal frame is windowed using triangular window defined as:

$$\omega(n) = \begin{cases} \frac{2n}{L}, 1 \le n \le \frac{L+1}{2} \\ \frac{2(L-n+1)}{L}, \frac{L}{2}+1 \le n \le L \end{cases} \quad (4)$$

where L- is the size of the window and n - is the sample number. This type of window ensures a higher accuracy of vowels detection than other shapes.

Vowel detection requires the initialization step which is performed in parallel to the initialization of the voice activity detection algorithm. In this step the threshold for the *PVD* is calculated as the mean value of first 40 frames of the signal according to the formula (Eq. 5):

$$Pth = C \frac{\sum_{n=1}^{N} PVD(n)}{N} \quad (5)$$

where *Pth*- is initial value of the threshold, $PVD(n)$ - is the value of peak-valley difference for the n-th signal frame, N - is number of frames that were used for initial threshold calculation, C - is correction factor. The correction factor was selected experimentally and was set to 1,1.

For every signal frame *PVD* value is determined and smoothed by calculating the average of last three values. The signal frame is marked as a vowel when: the value of the smoothed *PVD* is higher than *Pth* threshold and it has a local maximum in the *PVD* curve or its value is higher than 75 % of the value of the last local maximum. If the value is lower than *Pth* then the decision of voice activity detector is corrected and frame is marked as silence. For other situations frame is assigned to the consonant class.

In the real-time analysis assumptions presented above are tested in the following manner:

1. if the *PVD* for the frame *n-1* is greater than for frames *n* and *n-2,* where *n* is the number of the current analysis frame, and greater than *Pth* threshold, then frame *n-2* is marked as vowel and information about peak detection in *n-1* frame is saved,
2. if the condition 1 is not fulfilled, the second condition is checked, namely: if the information about peak presence is up to date and *PVD* value for the frame *n-2* is greater than 75 % of that peak, then frame *n-2* is marked as vowel,
3. if conditions 1 and 2 are not fulfilled and *PVD* value for the frame *n-2* is lower than *Pth,* then decision obtained using voice detector is corrected and frame is marked as noise (information about peak presence is canceled),
4. otherwise frame *n-2* is marked as consonant and information about peak presence is canceled.

An example of the vowels detection in real-time conditions is presented in Fig 2.

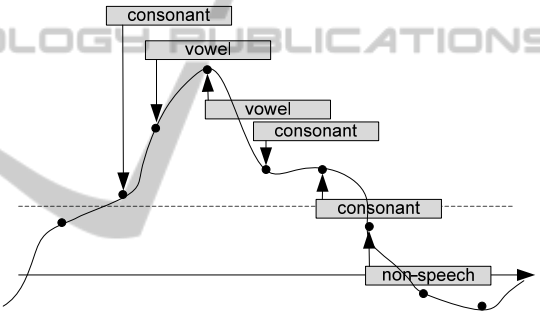

Figure 2: Vowels detection in real-time. Grey boxes represent analysis frames, dotted line represents *Pth* value.

Despite the fact that a simple voice activity detection algorithm was used, false positive errors appearing during the classification, have no impact on the vowel detection algorithm. It is achieved owing to the third step of the vowel detection algorithm, where all misclassified noise frames are detected and removed from the analysis.

2.3 ROS Estimation

ROS is a useful parameter in many speech processing systems. For the most part it is used in the automatic speech recognition (ASR). In the ASR many speech parameters are highly related to ROS. Hence, ROS is used to adjust the HMM model for different speech rates (Narayanan, 2005).

In the literature several definitions of ROS can be found. All of them require speech signal segmentation. For example Narayanan defines ROS as a number of syllables per second (SPS). In other

works ROS is defined inter alia as: phones per second (PPS) (Mirghafori, 1996), vowels per second (VPS) (Pfau, 1998), phones per second normalized to the probability of the specific phone duration (Zheng, Franco, Stolcke, 2000), word duration normalized to the probability of its duration (Zheng, Franco, Weng, 2000). Some measures, like those proposed by the Zheng et al., required the ASR or the transcription of the utterances. Therefore, for real-time unknown input signal, ROS estimation could be done only by statistical analysis. In this work, as ROS definition, the VPS parameter is used, as the derivate of SPS measure. Therefore, ROS is defined as (Eq. 6):

$$ROS(n) = \frac{N_{vowels}}{\Delta t} \quad (6)$$

For every signal frame ROS estimation is performed using the knowledge about the frame content, which is provided by vowels and voice activity detectors. Therefore, ROS value is updated for every 23 ms (length of the vowel detector analysis frame). Instantaneous ROS value is calculated as the mean number of vowels in the last 2 s of speech signal. Period of the time for the averaging was chosen experimentally in such a way that local ROS changes could be captured.

The highest ROS value that could be measured by this method equals 21 vowels/s, provided that all vowels and consonants durations are equal to 23 ms. It is worth mentioning that the instantaneous value of ROS is updated only when the current frame does not contain silence or prolongation of the vowel. At the beginning of the algorithm work, to eliminate the situation when the ROS values increase from zero to some value, initial ROS value is set to 5,16 vowels/s.

During the analysis instantaneous ROS value is used to assign, to the current utterance, one of speech rates categories, high or low. This division is obtained using the ROS threshold value (*ROSth*). *ROSth* was determined during the analysis of the mean ROS values of the speech rates recorded for 8 persons. Each person read five different phrases with three speech rates: high, medium and low. Results of the ROS statistics were presented in Tab. 1.

Table 1: Mean value and standard deviation of ROS calculated for the different speech rates.

speech rate	low	medium	high
μ(ROS)[vowels/s]	4,80	5,17	5,52
σ(ROS)[vowels/s]	0,76	0,75	0,79

It can be seen that, because of the high value of the standard deviation (nearly 0.76 for all classes) and as a consequence of the low distance between the neighbouring classes, only two classes could be separated linearly using the instantaneous ROS value. On the basis of the statistics, the ROS value was set to 5.16 vowels/s. The threshold was calculated according to the equation (7):

$$ROS_{th} = \frac{\mu(ROS)_{low} + \mu(ROS)_{high}}{2} \quad (7)$$

where $\mu(ROS)_{low}$ is the mean value of ROS for the low rate speech and $\mu(ROS)_{high}$ is the mean value of ROS for the high rate speech.

In Sec. 3 the accuracy of speech rate class recognition as well as its applicability to the non-uniform speech stretching are investigated.

2.4 Time-scale Modification Algorithm Selection

Many algorithms dedicated for speech time-scaling can be found in literature. All of them are based on the overlap-and-add technique. Most of the known algorithms were not optimized for real-time signal processing. Therefore, for real-time speech stretching only a few methods could be used. The best quality of time-scaled speech is achieved for complex methods that combine precise speech signal analysis such as speech periodicity judgment and adjustment of the analysis and synthesis frame sizes to the current part of the signal (Moulines, 1995). The algorithms, for instance PSOLA (Pitch Synchronous Overlap and Add) or WSOLA (Waveform Similarity Based Overlap and Add) produce high quality modified signals (Grofit, 2008; Verhelst, 1993), but require changing analysis shift sizes (WSOLA) or synthesis (PSOLA) frame sizes according to the current speech content.

It was shown that those algorithms could be used for real-time signal processing (Verhelst, 1993; Le Beux 2010), but for the non-uniform time-scale modification variable sizes of analysis time shift or synthesis frame would add complexity to the detection algorithms (voice activity detection, vowels detection). For this reason, NU-RTSM algorithm is based on the SOLA algorithm (Synchronous Overlap-and-Add) which in the fundamental form uses constant values of the analysis/synthesis frame sizes and analysis/synthesis time shift (Pesce, 2000) as well ensures quality of the processed speech nearly as good as for the other methods (Verhelst; 1993; Kupryjanow, 2009).

2.5 SOLA based Non-uniform Time-scale Modification Algorithm

To achieve a high quality of the stretched speech, analysis/synthesis frame size and analysis time shift should be selected properly i.e. frame length *L* should cover at least one period of the lowest speech component and in the synthesis stage, for all used scaling factors α(t), overlap size should be at least *L*/3 length. For the designed algorithm *L* value was set to 46 ms and analysis time shift *Sa* to 11,5 ms.

The synthesis time shift *Ss* is dependent on the current value of the scaling factor α(t). The scaling factor is defined as:

$$\alpha(t) = \frac{Ss}{Sa} \tag{8}$$

Synchronization between two synthesized overlapped frames is obtained by calculating the highest similarity point which is determined by the maximum of the cross-correlation function calculated for the overlapped parts of successive frames.

To reduce the duration of the stretched speech and to improve the quality of the modified signal, the scaling factor is changed for different speech content. For both speech rates (low and high) vowels are stretched up with the designed scale factor value ($\alpha(t)=\alpha_d$, being the value that is specified for the processing), and noise is not modified (α(t)=1) or removed from the signal dependently on the input/output synchronization state. For the low rate speech consonants are stretched up with the factor lower than α_d and equal to $\alpha(t)=0{,}8\cdot\alpha_d$, and for the high rate speech consonants are not stretched (α(t)=1). As it was shown in the third Sec. of this paper, the quality of speech preserved with the proposed method is better than the quality achieved with typical uniform SOLA algorithm.

3 EXPERIMENTS

The evaluation of the proposed algorithms was presented in this section. All algorithms were implemented in Matlab in such a way that the real-time signal processing was simulated. Sampling frequency of processed signals was set to 22,05 kHz.

3.1 Rate of Speech Estimation

The proposed method of real-time ROS estimation was tested using 80 recordings of 8 persons (1 woman, 7 men). Each person spoke five different phrases with 2 different speech rates: low, and high. Tab. 2 presents the accuracy of the speech rate detection. It can be seen that for the slow speech rate nearly 73 % frames were recognized correctly and for the high rate speech: 66 %. The main errors are connected to the fact that the rate of speech in the recording was not ideally constant.

In Fig. 3 waveforms corresponding to the recorded male high rate speech with the detected vowels and estimated speech rate are presented. It can be observed that the main error occurs at the beginning of the ROS extraction. It is connected to the fact that the ROS algorithm assumes that the most probable is low speech rate as a typical speech rate of every person (it is assumed in the ROS initialization phase). The second type of error can be seen after the prolongation of the vowel. It is connected to the fact that the current value of ROS is highly related to the historical data, so ROS estimation needs several new frames to enable following high variations of the instantaneous ROS.

Table 2: Percentage number of frames marked as low/high rate speech, calculated for female and male speech expressed with low and high rate.

Speech rate	low rate speech recording	high rate speech recording
Low	**72,67**	34,15
high	27,32	**65,84**

3.2 Time-scaled Speech Assessment

Quality of NU-RTSM algorithm was assessed in subjective tests performed for 19 healthy persons (2 women, 17 men). Each person had to assess the quality of the speech stretched using the typical SOLA algorithm implementation and the proposed NU-RTSM algorithm. Two values of the stretching factors were chosen: 1,9 and 2,1. Four recordings were used during the experiment: two spoken with the low rate, and two with the high rate. Both of them were spoken by a woman and a man. In all recordings the same phrase was uttered.

Three parameters were rated during tests: signal quality, speech naturalness and speech intelligibility. The assessment was made using the following scale: 1- very poor, 2- poor, 3-medium, 4-good, 5-very good. In Figs. 4-6 histograms of the speech stretching assessment are presented. It can be seen that for both speech rates, as well as for all parameter values, histograms that represent NU-RTSM assessment have higher placed gravity centres than for the SOLA algorithm. For the high rate speech this difference becomes more significant.

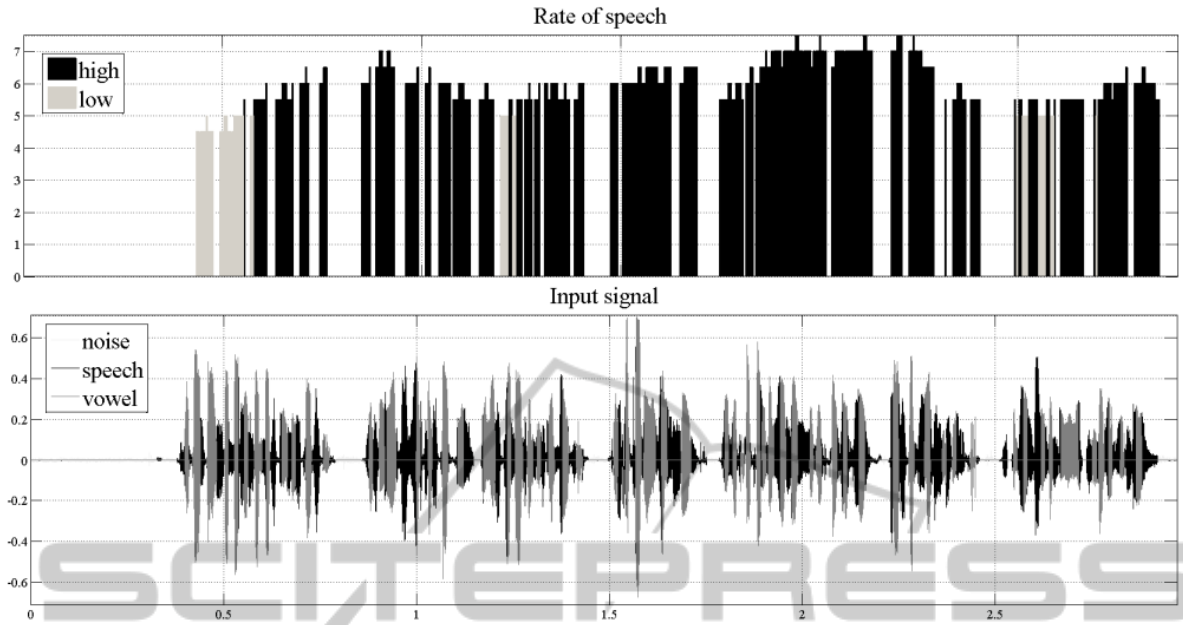

Figure 3: Speech rate recognition for high rate male speech.

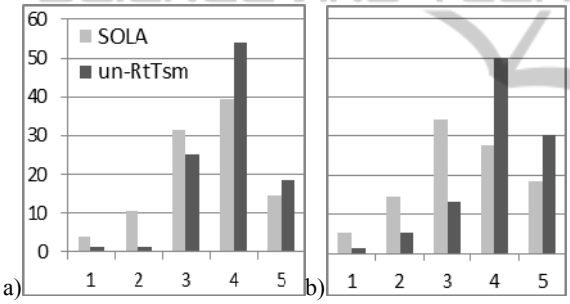

Figure 4: Signal quality assessment for different speech rates: a) low, b) high.

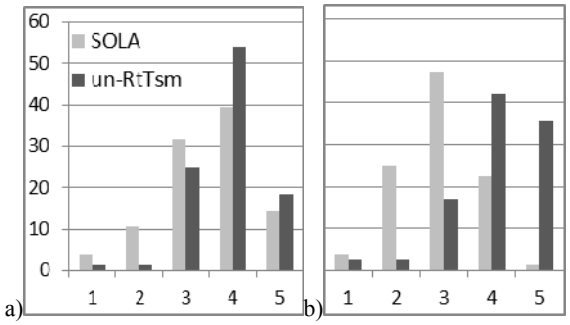

Figure 5: Speech naturalness assessment for different speech rates: a) low, b) high.

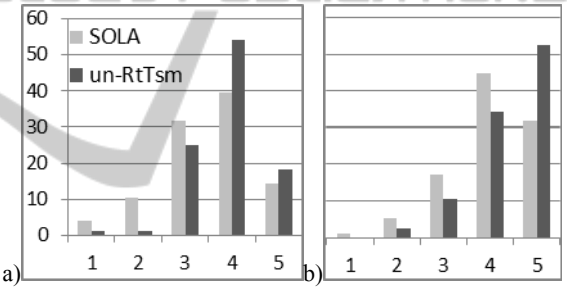

Figure 6: Speech intelligibility assessment for different speech rates: a) low, b) high.

4 CONCLUSIONS

The proposed Non-Uniform Speech Real-Time Speech Modification algorithm ensures high quality of the stretched speech. Subjective tests have shown that naturalness and intelligibility of the processed speech is higher than in case of a typical uniform signal stretching. In the future implementation of the algorithm real-time mode should be enabled on a mobile device. Moreover, speech perception tests for the people with hearing time-resolution problems should be made in order to verify modification usability.

ACKNOWLEDGEMENTS

Research funded within the project No.POIG.

01.03.01-22-017/08, entitled "Elaboration of a series of multimodal interfaces and their implementation to educational, medical, security and industrial applications". The project is subsidized from the European Regional Development Fund by the Polish State budget".

REFERENCES

Demol, M., Verhelst W., Struye K., Verhoeve P., 2005. Efficient Non-Uniform Time-Scaling of Speech with WSOLA. *Speech and Computers (SPECOM).*

Grofit, S., Lavner, Y., Jan. 2008. Time-Scale Modification of Audio Signals Using Enhanced WSOLA With Management of Transients.*IEEE Trans. On audio, speech, and language processing, vol. 16, no. 1.*

Kupryjanow, A., Czyzewski, A., London, 2010. Real-time speech-rate modification experiments. *Audio Engineering Society Convention Paper, preprint No. 8052.*

Kupryjanow, A., Czyzewski, A., Poznań 2009. Time-scale modification of speech signals for supporting hearing impaired schoolchildren. *Proc. of the International Conference NTAV/SPA, New Trends in Audio and Video, Signal Processing: Algorithms, Architectures, Arrangements and Applications, pp. 159-162.*

Le Beux, S., Doval, B., d'Alessandro, C., 2010. Issues and solutions related to real-time TD-PSOLA implementation. *Audio Engineering Society Convention Paper, Preprint No. 8085.*

Mirghafori, N., Fosler, E., Morgan, N. 1996. Towards Robustness to Fast Speech in ASR. *Proc. ICASSP'96, pp. I335-338.*

Moattar, M., Homayounpour, M., Kalantari, N., 2010. A new approach for robust realtime voice activity detection using spectral pattern. *ICASSP.*

Morgan, N., Fosler-Lussier ,E., Seattle, 1998. Combining multiple estimators of speaking rate.Seattle. *ICASSP.*

Moulines, E., Laroche, J., 1995. Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication 16(2): 175-205.*

Narayanan, S., Wang D., 2005. Speech rate estimation via temporal correlation andselected sub-band correlation. *ICASSP.*

Pesce, F., Italy, 2000. Realtime-stretching of speech signals. *DAFX.*

Pfau, T., Ruske, G., 1998. Estimating the speaking rate by vowel detection. *IEEE.*

Tallal ,P. et.al, 5 January, 1996. Language Comprehension in Language-Learning Impaired Children Improved with acoustically modified speech.*Science, Vol. 271.*

Verhelst, W., Roelands, M., 1993. *An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech.*

Zheng, J., Franco, H., Stolcke, A., 2000. *Rate of Speech Modeling for Large Vocabulary Conversational Speech Recognition.*

Zheng, J., Franco, H., Weng, F., Sankar, A., Bratt, H., 2000.Word-level rate-of-speech modeling using rate-specificphones and pronunciations. *In: Proc. IEEE Int. Conf.Acoust. Speech Signal Process, Istanbul, Vol. 3, pp. 1775–1778.*