A Method of Real-Time Non-uniform Speech Stretching

Adam Kupryjanow and Andrzej Czyzewski

Multimedia Systems Department, Gdansk University of Technology
Narutowicza 11/12, Gdansk, Poland
{adamq,andcz}@sound.eti.pg.gda.pl

Abstract. Developed method of real-time non-uniform speech stretching is presented. The proposed solution is based on the well-known SOLA algorithm (Synchronous Overlap and Add). Non-uniform time-scale modification is achieved by the adjustment of time scaling factor values in accordance with the signal content. Dependently on the speech unit (vowels/consonants), instantaneous rate of speech (ROS), and speech signal presence, values of the scaling factor are selected. This provides as low as possible difference in the duration of the input and output signal and high naturalness and quality of the modified speech. In the experimental part of the paper accuracy of the proposed ROS estimator is examined. Quality of the speech stretched using the proposed method is assessed in the subjective tests.

Keywords: Time-scale Modification, Voice Detection, Vowels Detection, Rate of Speech Estimation.

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 [8], reduction of computational complexity or its adaptation for real-time signal processing [10]. 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 [12]. 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 [3]. Some improvements of that algorithm are proposed, i.e. usage of non-uniform time-scaling, in this paper.

As it was shown by Demol [1], 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

M.S. Obaidat, J.L. Sevillano, and J. Filipe (Eds.): ICETE 2011, CCIS 314, pp. 362–373, 2012.

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 combination of voice activity detection, vowels detection, rate of speech estimation and time-scale modification algorithms.

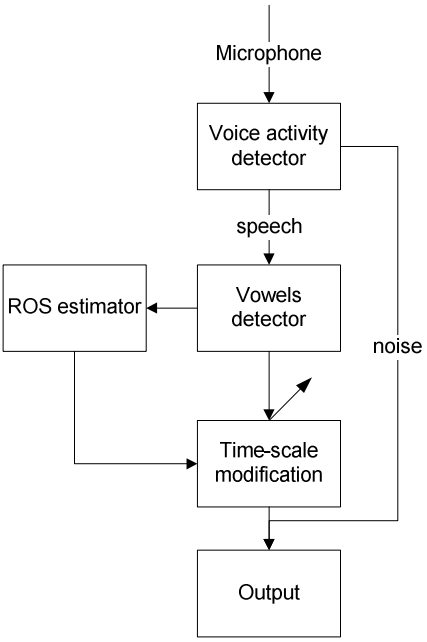

Fig. 1. NU-RTSM block diagram

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,