Linear predictive based voice transformation module for Macedonian TTS

Branislav Gerazov, *Graduate Student Member, IEEE*, Sofija Bogdanova, *Senior Member, IEEE*, and Zoran Ivanovski, *Member, IEEE*

Abstract **– The paper presents the implementation of a voice transformation module in the Macedonian text-to-speech system “Speak Macedonian”. The voice transformation algorithm is based on changes of speech formant structure through the compression or expansion of the frequency response of the vocal tract filter obtained with linear predictive analysis. Average pitch modification and “breathiness“ are included . The module expands the functionality of the system, allowing simulation of different voices with the standard voice’s unit inventory.**

Key Words**– formant, linear predicitive analysis, pitch, source-filter model, text-to-speech synthesis, voice transformation.**

I. Introduction

Concatenative synthesis is the most common paradigm used in text-to-speech (TTS) systems of today. It is based on concatenation of prerecorded segments of natural speech to generate the requested speech output. This approach gives the synthetic speech a very natural sound, without the burdensome need to build a model of the speech production process.

The prerecorded segments are stored in a data base. The larger the database the better the quality of the synthesized speech. For example high-end unit-selection TTS systems usually contain hours of recorded speech material with many representations of the same unit. The most extreme example being Japanese XIMERA, with a 170 hour, 25,5 GB database of recorded speech, [1].

The creation of such long segment (unit) databases is a time-consuming task. It involves the realization of long controlled recording sessions with the speaker who’s voice is to be included in the system, the segmentation of the recorded speech signal into the needed speech segments, and the phonetic labeling of the segments. Speech segmentation and phonetic labeling are extremely time-consuming endeavors, especially when the recorded speech length is measured in hours. Even with tools for automatic speech segmentation/labeling at hand, a final human expert verification is always needed. Because of this man-hour cost associated with the development of unit inventories, concatenative TTS systems usually come with only a couple of voices. This is where voice transformation algorithms come into play.

Voice transformation (VT) can alleviate the need of recording additional voices by reusing the already created voices to generate new and different ones. Voice transformation algorithms seek to alter the speech signal in such a way that the output speech seems to be spoken by a different person than the original speaker, while maintaining full intelligibility and maximal naturalness. This is usually done by modeling the speech signal and then changing the parameters of the model at hand to resynthesize the speech, [2].

The VT algorithm presented in this paper is a simple but effective algorithm based on the powerful source-filter model of human speech production. The algorithm alters formant frequencies and bandwidths through modification of the linear prediction (LP) modeled spectrum of the synthesis filter transfer function. The algorithm also changes the average pitch of the speech signal through the implemented PSOLA (Pitch-Synchronous Overlap and Add) algorithm. The VT algorithm is integrated as a VT module in the Macedonian TTS system “Speak Macedonian” developed at the Faculty of Electrical Engineering and Information Technologies (FEEIT), [3].

II. The source-filter model and LP basics

The source-filter model is one of the most powerful concepts used in the coding, analysis and synthesis of digital speech. The model represents speech generation with a system comprising an excitation source and a synthesis filter which roughly correspond to the vocal cords and the vocal tract, respectively, Fig. 1.

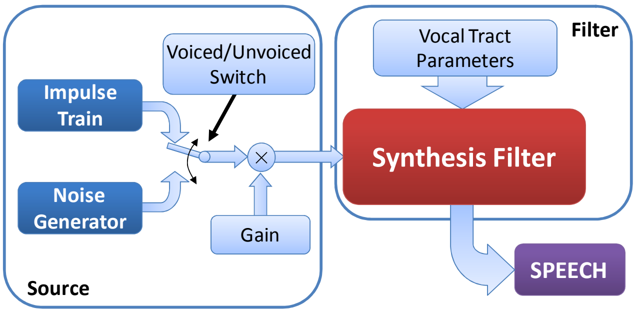

Fig. 1 – The source-filter speech production model

Branislav Gerazov, Sofija Bogdanova, and Zoran Ivanovski, Faculty of Electrical Engineering and Information Technologies, Ruger Boskovik, PO Box 574 - Skopje, Macedonia (tel.: +389 2 3099 191; fax: +389 2 3064 262; e-mail: {gerazov, sofija, mars @feit.ukim.edu.mk}).

The source model, in its most rudimentary form, is comprised of two generators. The first generates an impulse train, mimicking the pulse excitation generated by the vocal cords for voiced speech. The second generates white Gaussian noise, mimicking the noise-like excitation used by humans to generate unvoiced speech. It also includes a switch to chose between the two and a Gain factor control.

The synthesis filter approximates the transfer function of the vocal tract modeling it with an autoregressive model. Its role is to shape the signal generated by the source and give it phonetic character. The filter parameters are extracted from the speech input by the use of linear predictive (LP) analysis, [4].

This simple model was used in the first LP based codec's such as LPC-10 (Linear Predictive Coding 10), also known as FS-1015 (Federal Standard 1015), [5]. The codec gives a large compression ratio, however with terrible naturalness, due to the use of an oversimplified source model. Further development of LP-based codec's, most notably the CELP (Code-Exited Linear Prediction) codec used in GSM telephony, overcame this by using more sophisticated models for the source. For example in CELP the source is modeled using a number of different excitation sequences stored in an excitation codebook, [6].

In LP-based voice transformation we are only concerned with the synthesis filter. Its transfer function of the synthesis filter is given by

$$H(z) = \frac{S(z)}{E(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad (1)$$

where G is the filter gain, a_k are the filter coefficients and p is the filter order.

The speech signal is thus modeled as:

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + G \cdot e[n] \qquad (2)$$

where $s[n]$ is the speech signal and $e[n]$ is the excitation signal.

In terms of prediction, a linear predictor with prediction coefficients α_k is given by:

$$\tilde{s}[n] = \sum_{k=1}^{p} \alpha_k s[n-k] \qquad (3)$$

where $\tilde{s}[n]$ is the predicted current sample of $s[n]$ based on p previous samples. Thus the prediction error is

$$\varepsilon[n] = s[n] - \tilde{s}[n] = s[n] - \sum_{k=1}^{p} \alpha_k s[n-k] \qquad (4)$$

from (2) we finaly have

$$\varepsilon[n] = G \cdot e[n] \qquad (5)$$

that holds when $\alpha_k = a_k$.

Thus we have separated the excitation signal from the filter and can later on use it to resynthesize the speech waveform, but with the modified synthesis filter.

III. LP-BASED VOICE TRANSFORMATION

The main idea in LP-based voice transformation is to transform the synthesis filter in such a way that formant frequencies shift to the left/right on the frequency axis, i.e. its transfer function compresses/expands, while leaving the excitation signal intact. Because the formant frequencies reflect the size and shape of the vocal tract their location change will have the effect of making the speech sound as spoken by a different person – the goal of voice transformation. Furthermore, because the formant frequencies need only be in a particular range of frequencies for the speech to be intelligible we have the freedom of shifting them quite a bit.

The algorithm takes the following steps:

1° Frames are extracted from the speech signal with a Hamming window with a time length of 40 ms at 10 ms steps.

2° For each frame, the filter coefficients are calculated recursively using the Levinson-Durbin algorithm. The filter order is set to p = 25 for the used sampling rate of 16 kHz, as recommended in [4], i.e. 2 poles for each kHz of the speech signal and 3-4 extra for modeling the zeros in the vocal tract transfer function due to nasals.

3° The excitation signal of the current frame is calculated using (4).

4° From the filter coefficients the transfer function of the synthesis filter is calculated.

5° The transfer function is then compressed or expanded in accordance to the value of a transformation coefficient k giving the new synthesis filter transfer function. The range that k can have is set from 0.5 to 2.

6° The new synthesis filter is used to filter the unaltered excitation signal extracted in step 3.

7° Output frames are reassembled with an overlap-add operation that yields the output speech signal.

The compression and expansion of the synthesis filter transfer function, not only changes the formant frequencies but also their respective bandwidths. This provides formant structure change simulating differences in the vocal tracts among speakers.

Use of the algorithm is depicted in Figs. 2 & 3 where a single frame containing the phone "a" is processed. Fig. 2 shows transfer function compression for k = 0.5 and 0.8, corresponding to compression to 50 and 20 %. Fig. 3 shows transfer function expansion for k = 1.4 and 2, corresponding to expansion to 140 and 200 %.

IV. AVERAGE PITCH MODIFICATION

In addition to changing the formant frequencies and bandwidths of the synthesis filter, an important part of the implemented voice transformation module is the average pitch modification functionality. The functionality is based on controlling the built-in PSOLA algorithm of the TTS system "Speak Macedonian" for generating the output speech prosody (intonation, accent, rhythm).

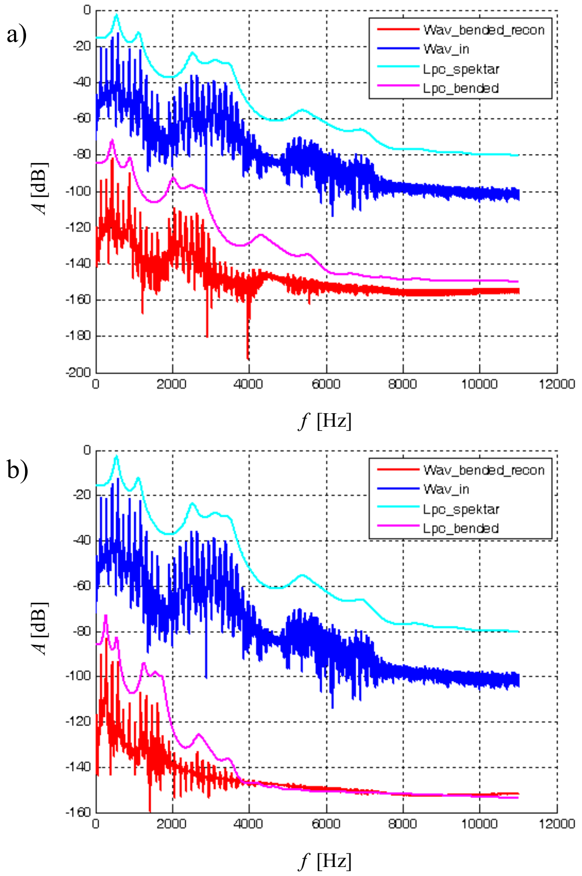

Fig 2 – Synthesis filter transfer function compression for: a) k = 0.5, and b) k = 0.8

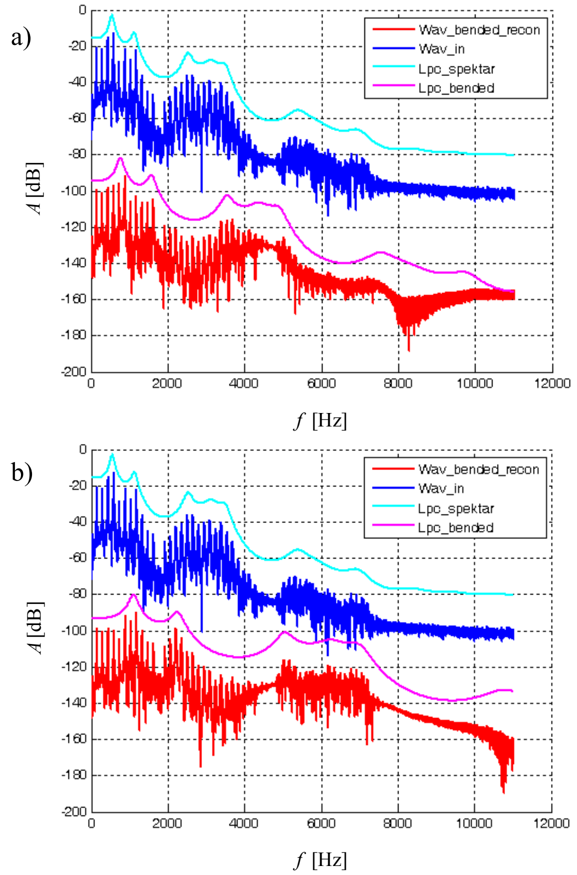

Fig 3 - Synthesis filter transfer function expansion for: a) k = 1.4, and b) k = 2

The average pitch range in speech can go from below 100Hz for low-pitched male speakers to over 250Hz for high-pitched voices of women and children. The standard f_average in our system is set to 150 Hz. The VT module allows change of this value from 50 to 300 Hz. In this way the differences in vocal folds and vocal tract length between speakers can be simulated, yielding easy differentiation into male/female/child voices.

V. Breathiness

The VT module also includes an option to give the output speech a "breathy quality" or "breathiness" which is useful when simulating elderly speakers. This is done by adding a noise signal to the extracted excitation signal before it is filtered by the transformed synthesis filter in step 6o in the algorithm. The degree of "breathiness" is controlled by the parameter b which designates the noise level in percent relative to the maximum of the excitation signal of the current frame. The VT module allows change of this value in the range from 0.1 to 2.

VI. Integration in the TTS System Interface

The voice transformation module is well integrated in the graphical user interface (GUI) of the "Speak Macedonian" TTS system giving the user easy and intuitive control, Fig. 4.

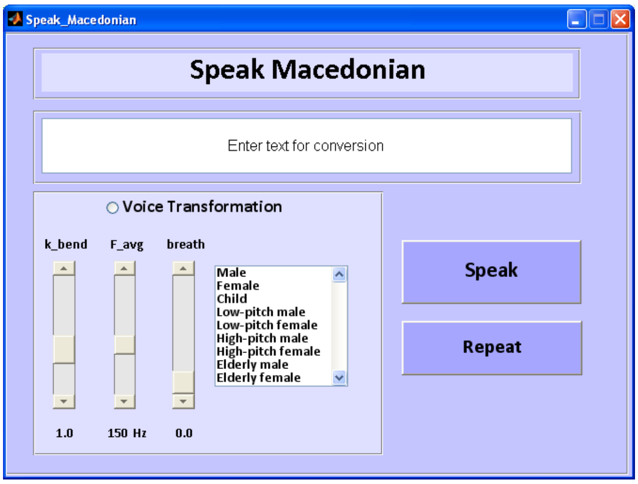

Fig. 4 – The "Speak Macedonian" GUI with voice transformation module controls

The voice transformation controls field consists of three sliders controlling:

- k – the synthesis filter transfer function compression/expansion coefficient
- F_avg – the average voice pitch
- breath – the amount of "breathiness" added to the transformed speech

Also included is a list of available voice transformation presets. The parameters that correspond to these presets were given in Table. 1. The values given here correspond to the standard male 150 Hz voice used in the system. Breathiness is only added to the elderly male and female VT presets with a value of b = 1.5.

VII. Results

The presented algorithm was tested for various parameter combinations. The degree of voice transformation that were obtained is substantial. At times it was hard to believe that the results were derived from the same speech signal. A total of 9 distinct voice transformations was selected as representative and included as presets for the standard male speaker voice of the system. Many more alterations can be obtained by the user himself, thus allowing him to fine tune the system's voice to his own liking.

Table 1: Average pitch and transformation coefficient k presets implemented in VT module

VT target	*f_average [Hz]*	*k*
Male	/	/
Female	200	1.2
Child	300	1.6
Low-pitch male	80	0.8
Low-pitch female	170	1.4
High-pitch male	200	1
High-pitch female	250	1.2
Elderly male	150	0.8
Elderly female	200	1.4

The computational complexity of the algorithm does not burden the speech synthesis process dramatically, because of its relative simplicity. The quality degradation of the transformed speech is minor and is relative to the parameters used in the voice transformation. Decrease of quality is more evident when extreme values for the parameters are chosen, and they should be avoided.

VIII. Conclusion

The presented voice transformation algorithm is based on the source-filter model of speech production. The algorithm changes the formant structure of the input speech through compression and expansion of the synthesis filter transfer function. Average pitch control is added to complete the voice transformation effect. As an addition a breathiness effect was also added to expand the functionality of the algorithm.

The algorithm was successfully added to the "Speak Macedonian" TTS system. It is not a computationally complex algorithm and thus had no major impact on the system's performance. The results of the voice transformation were more than satisfactory giving at least 9 distinct voices from the single standard male voice, implemented as presets in the module. The integration of the module provides an increased flexibility of the TTS system in providing a richer experience to the users.

Even more, the presented algorithm can be easily added to other TTS systems enhancing their own functionality.

References

[1] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, K. Tokuda, "XIMERA: A new TTS from ATR based on corpus-based technologies", *Proc. 5th ISCA Speech Synthesis Workshop 2004* pp. 179–184

[2] Stylianou Y. "Voice Transformation", in *Springer Handbook of Speech Processing*, Benesty J., Sondhi M. M., Huang Y. Eds., Springer-Verlag Berlin Heidelberg, 2008, pp. 489 - 502

[3] B. Gerazov, G. Shutinoski and G. Arsov, "A Novel Quasi-Diphone Inventory Approach to Text-To-Speech Synthesis", *MELECON '08*, Ajaccio, France, May 5-7, 2008

[4] Rabiner L.R., R.W. Schafer, "Digital Processing of Speech Signals", Prentice Hall, 1978

[5] Tremain T. E., "The Government Standard Linear Predictive Coding Algorithm, LPC-10", *Speech Technology*, April. 1982, pp. 40–49

[6] Chu W., "Speech Coding Algorithms - foundation and evolution of of standardized coders", Wiley, 2003