Phonetic-Acoustical Problems of Personal Voice Cloning by TTS

Boris M. Lobanov and Lilia I. Tsirulnik

United Institute of Informatics Problems, Nat. Ac. of Sc. Belarus
e-mail: Lobanov@newman.bas-net.byT

Abstract

The report describes a recent development of a TTS-system for Russian based on allophonic natural speech signal elements (about 1500 in all) with the maximal possible imitation of individual male and female voices. In distinction to the biological task of cloning, the target is not a copy of the human being as a whole but of only one of its functions, particularly, that of reading aloud an orthographically unrestricted text preserving thereby the individual acoustic characteristics of a speaker's voice, as well as his/her phonetic (segmental and prosodic) peculiarities. A successful solution of the task outlined above presupposes that the following two requirements should be unequivocally satisfied:

- The fullest possible use of a complex of acoustic characteristics carrying information about the individual voice and pronunciation properties of the speaker being imitated.
- The minimal possible distortions of the elements of concatenation at all stages of their 'production', as well as the maximal possible accuracy of prosodic modifications in the process of speech synthesis.

1. Introduction

The aim of preserving the speaker's personal phonetic and acoustical characteristics by TTS resembles, although distantly, the widely-known biological problem of cloning, whereby on the basis of a comparatively small amount of genetic information an attempt is made of reproducing a living being copy as a whole. In our case, an attempt is being made of creating a close copy, not a biological one, but a computerized one, and not of the whole human being, but of one of its intellectual functions only: the reading aloud of a piece of orthographic text. A task is hereby set of preserving, as fully as possible, the personal acoustic peculiarities of the voice, the phonetic peculiarities of the pronunciation of segmental sounds and the individual prosodic features, i.e. the individual peculiarities of the tonal, rhythmical and dynamic organization of speech. In principle, there exists in genetics a possibility of producing 'chimera'-like creatures from heterogeneous genetic material. As far as 'cloning' voices and speech is concerned - this is the case when synthesis is produced with the acoustics of the voice of one speaker, the phonetic features of sound articulation of another, and the prosodic characteristics of a third one.

The aim of preserving the speaker's personal phonetic and acoustical characteristics is made feasible, first of all, by utilizing fragments of natural speech signal (sound waves coextensive with allophones, in our case). An important factor here is also using a sufficiently wide inventory of allophones covering all of the most significant positional and combinatory variations of phonemes. The speaker's individual prosodic characteristics are captured by identifying and implementing the features that are most essential for the prosodic organization of speech in principle and for the prosodic 'portrait' of the given speaker, in particular.

Minimization of distortions in the process of element concatenation is achieved by means of the high-quality digitizing of the relevant acoustic signals as well as precision of marking the boundaries between the allophones and their pitches (periods). An obligatory requirement here is avoiding, as much as possible, any kind of additional transformations of the recorded signals in the process of imparting prosodic modifications like the PSOLA [1] or FFT [2], which cause inevitable approximation errors. Instead, special ***"****splicing"* algorithms are proposed of sound-wave modifications in the course of F_0 changes. According to these algorithms the natural sound wave remains unchanged on one part of the period, that corresponds the time when vocal folds are closed, while on the other part the wave patterning is approximated or predicted in some way.

2. General Structure of the Synthesizer

Synthesis of phonetic characteristics of speech is based on the allophone-wave method of speech signal concatenation. The basic principle of synthesizing the prosodic characteristics of speech is the division of an utterance into accentual groups and the formation on their basis of entire tonal, rhythmical and dynamic contours of a syntagm and utterance as a whole [3,4].

General structure of the synthesizer is shown in fig. 1. The incoming orthographic text undergoes a number of successive analytical operations carried out with the help of specialized processors.

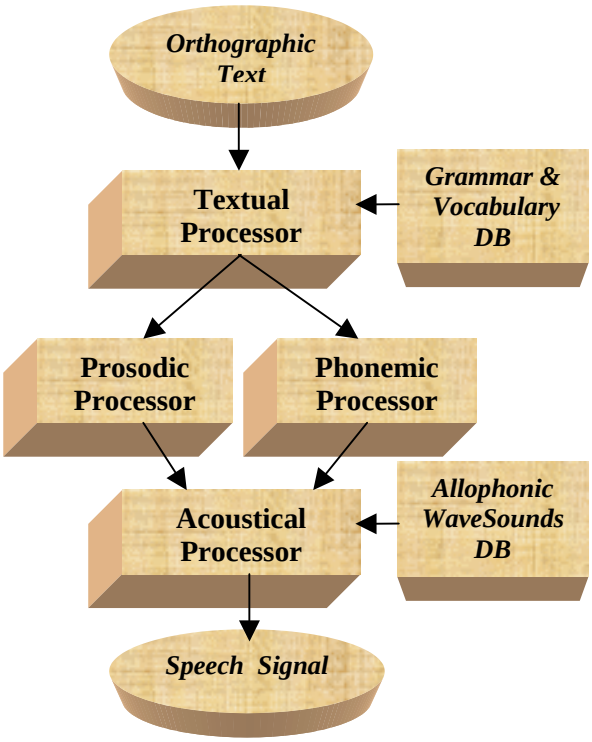

Fig. 1. General structure of the synthesizer

A *textual* processor is devised to transform the incoming orthographic text into a marked phonemic one. The processor also performs the tasks of placing wordstress and marking intonation within the syntagms. The marked phonemic text is then fed to two processors: prosodic and phonetic. In the *prosodic* processor the phonemic text is divided up into Accentual Units (AU) which are further split into their constituent elements: prenuclear, nuclear and postnuclear parts. And, finally, the prosodic processor fulfils the functions of determining the values of the amplitude (A), the phoneme duration (T) and the fundamental frequency (F0) for each Element of the Accentual Units (EAU). The *phonetic* processor generats positional and combinatory allophones from the incoming phonemic text. The *acoustic* processor uses the information as to which allophones it is necessary to synthesize, as well as which prosodic characteristics should be ascribed to each allophone and generates the speech signal by concatenating portions of allophone sound waves and their modifications in accordance with the required current values of F0, A, T.

The textual processor is the most universal block, whose structure and functioning display the least dependence on the individual features of the speaker being imitated. However, here too, there is a possibility of modifications connected with the individual peculiarities of text segmentation on the syntagms. This is illustrated in fig.2 where the probability of the syntagme presence with one, two, three or four AU for three diferrent speakers but for the same text.

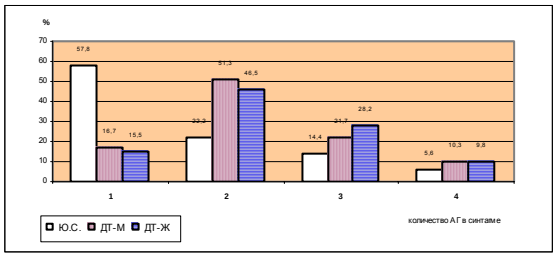

Fig.2. The probability of the one, two, three or four AU presence in syntagme for three diferrent speakers

A considerably larger amount of information concerning the individual voice and speech characteristics is generated by the prosodic, phonetic and acoustic processors. Peculiarities of their structure and functions are considered below.

3. Technology of 'cloning' the acoustic characteristics of the voice

The personal acoustic characteristics of a speaker's voice are shaped by a number of factors, such as the structure and functioning of the speech organs (the larynx, the vocal folds, the pharynx, the mouth cavity and oth.), the dynamic peculiarities of the interaction of the vocal folds vibrations and the speech resonators (the coupling effect), and many others. Attempts at imitating personal voice characteristics by modeling the articulation and acoustic processes of speech production, have failed so far to bring about any noticeable results, primarily due to the extreme complexity of those processes. In this connection a more reasonable approach seems to consist in using fragments of natural speech waves as the minimal "genetic material" for 'cloning' a person's voice. It seems justifiable, too, to identify such a fragment with an allophone as the best studied phonetic entity. One can also assume that a limited set of allophones will ensure the possibility of generating oral speech of unrestricted content. Significantly, the sound wave contains all the personal peculiarities of voice production as they manifest themselves in a given concrete allophone.

The goal of 'cloning' the acoustic characteristics of a human voice requires that a database of allophonic sound waves should be created either from a sound corpus of a reasonable size, recorded by the given speaker or by using a sufficiently wide choice of the speaker's recordings on the radio or TV. The results discussed in the present paper have been obtained from the recordings of a specially designed experimental corpus, in which the number of selected words corresponds to the number of allophones being used. Each word has been selected on the criterion of the best possible representation of the given allophone. Principles of the allophone inventory solutions are discussed below in the paragraph dealing

with the 'cloning' of personal phonetic (segmental) features of pronunciation.

The recorded sound corpus is then analyzed by an expert with the help of a set of computer devices (both standard and original) used for processing the speech signals. The ultimate goal of this analysis is the creation of a sound-wave allophonic database. The database obtained in this way is stored in the form of signals in the Wav-format with the frequency of discretization of 22 kHz and the rating of 16 bites. Each Wav-file is supplied with a headline, which contains, in particular:

– the name of the allophone (three symbols, e.g. A132);
– the number of the signal samples - N;
– the number of pitches (periods) - K;
– the position of each pitch - P1, P2, ... Pk...Pk;
– the position of the medial pitch of an allophone - Ps;
– the amplitude of an allophone - A.

The number and position of each pitch in the order of a signal for each allophone are defined automatically with the help of a specially designed program - PITCH, based on the autocorrelational method of analyzing the periodicity of a signal. The position of a pitch is defined on the part of the period, that corresponds to the time when vocal folds are kept apart. The remaining 2 parameters - position of an allophone medial pitch - Ps and the amplitude A - are defined automatically.

The parameters of an allophone: Pk, Ps, A - are used in the process of synthetic speech generation for the prosodic modifications of of the fundamental frequency (F0), duration and intensity of a sound, respectively. The modification of F0 is obtained by means of modifying the duration of the current period of an allophone sound wave: by reducing it at the F0 increase and lengthening it when the F0 decreases. Modification of allophone duration is achieved by adding or removing the required number of signal periods in the position of a medial pitch - Ps. Modification of the sound intensity is achieved through a corresponding change of the signal amplitude.

As was mentioned above, the great harm to the voice personal acoustic characteristics is done by the wrong choice of the F0 modification technique, since it affects each period of a signal. That is why the strategy of synthetic speech personalization demands that special ‘sparing’ techniques of prosodic modification be worked out. It is necessary to preserve the greatest possible amount of information about the voice individual characteristics contained in the original speech wave and, at the same time, not to add the signal distortions of various kinds. Several methods were proposed and tested for direct F0 modification in time-domain representation of a speech signal:

- The method of local smoothing,
- The method of formants dumping,
- The method of local extension/compression,
- The method of smooth linear conjugation,
- The method of linear time-domain *“splicing”*.

As it has been found out, the best result is obtained by using the last of the above methods when the time-interval of splicing is chosen by 50% of the signal period for sound-wave modifications in the course of F_0 changes. According to this, the natural sound wave remains unchanged on the part of the period that corresponds to the time when vocal folds are closed.

The principle of splicing is generally the same as the one used for microwaves concatenation proposed in [4]. An example of the waves splicing is shown in fig.3.

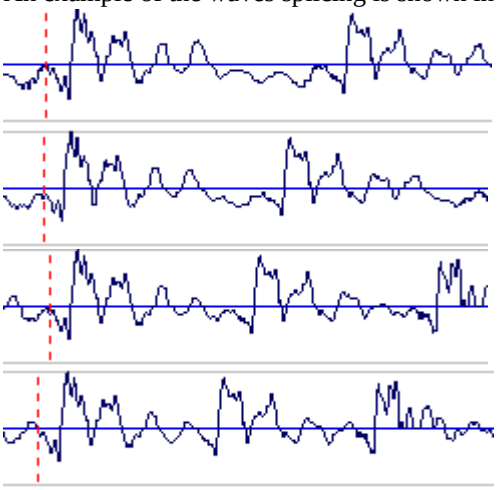

Fig.3. An example of waves splicing for the sound /A/. (Top-down: F0=80,100,120,140Hz)

4. 'Cloning' of the personal phonemic peculiarities of pronunciation

In distinction to voice personal acoustic characteristics, determined mainly by the static acoustic parameters of the vocal tract, the phonetic peculiarities of pronunciation are conditioned for the main part by the dynamics of articulatory motions, performed in the process of speech production. The rate of articulatory movements inherent to a given the individual, the characteristic delays or, vice versa, the speeding up of the movements of articulators, individual peculiarities of the articulation of this or that sound (ex. Russian /R/), a regional or a foreign accent and so on lead to the appearance of specific positional and combinatory variants of phonemes and constitute a unique system of allophones. Thus, it is possible to claim that successful 'cloning' of individual phonetic peculiarities depends, for the main part, on the successful imitation of the peculiarities of phoneme-allophone transformation inherent to a given speaker.

The method of phoneme-allophone transformation, proposed in this paper, ensures the generation of the following Russian vowel allophones: stressed (0), immediately pre-stressed (1), not immediately pre-stressed (2), after-stressed (3). There are 4 positions in all.

According to the left environment the following combinatory allophones of vowels are generated: after a syntagmatic pause (0), after most of the non-palatalised forelingual (1), labial (2) and velar consonants (3), then after /L/ (4), after /R/ (5), after /M/ (6), after /N/ (7), after most of the palatalised consonants (8), after /L'/ (9), after /R'/ (10), after /M'/ (11), after /N'/ (12), and after the vowels /U/ (13), /O/ (14), /A/ (15), /Y/ (17), /I/ (18). Total: 19 left contexts. Depending on the right context the following combinatory vowel allophones are generated. Before a syntagmatic pause (0), before forelingual and velar non-palatalized consonants and vowels /A/, /E/ (1); before labial non-palatalized consonants and the vowels /U/, /O/ (2); before forelingual and velar palatalized consonants and the vowel /I/ (4); before labial palatalized consonants and the vowel /Y/ (5). Total: 6 right contexts.

The final result in a general case is the generation of Nv = 4*19*6*6 (number of vowels) = 2736 Russian vowel allophones. Considering the well known positional and combinatory restrictions the actual number of allophones used in the synthesizer is less than 1500.

The generation of consonant allophones also takes account of the left and right context. The left context: after a pause (0), after voiceless consonants (1), after voiced consonants (2), after vowels (3). The right context: before a pause (0), before voiceless consonants (1), before voiced consonants (2), before unstressed (3) and befor stressed vowels (4). Thus, in a general case, the generation Nc = 4*5*36 (number of Russian consonants) = 720 consonant allophones is ensured. The total number of consonant allophones actually used in the synthesizer is less than 500.

5. 'Cloning' personal prosodic features of speech

The complex of prosodic parameters of speech, including melody, rhythm and prominence is provided by specific modifications of the fundamental frequency of voice - F0, duration of sounds - T, and the amplitude of sound signals - A. The main cause of prosodic modifications are regulated by the specific type of intonation utilized in a certain syntagme of the reading text, such as the intonation of completeness, incompleteness, interrogation, and exclamation. The kind of prosodic modifications is determined also by a number of other factors such as the type of text (prose, verse or conversation), the stylistic variety of speech (dictation, report, actor's reading aloud of a literary text). At this stage we are restricting ourselves with modeling narrative texts of a formal-neutral style (reading aloud of a report, a message, an instruction or any other kind of information).

The input symbols for intoning a text are the following: {new line} - end of a paragraph, {.} - full stop, {;} - semicolon, {:} - colon, {,} - comma, {-} dash, {(} - the beginning of a parenthetical word or a group of words, {)} - end of a parenthetical word or a group of words, {?} - interrogation mark, {!} - exclamatory mark. These symbols alongside of some of the conjunctions and other parts of speech serve as signals predicting the choice of an appropriate variant of the prosodic contour withing the categories of completeness, incompleteness, interrogation and exclamation.

In accordance with the principles of modeling intonation adopted in this work, the minimal prosodic unit is the Accentual Unit (AU), consisting of the nucleus (the fully stressed syllable), the pre-stressed part and the post-stressed part [3,4]. An AU may consist of one or more words. A syntagm, in its turn, may contain one or more AUs. The tonal, rhythmical and dynamic contours of the syntagm present a sequence of the prosodic portraits of AUs constituting the syntagm. For each of the above variants of intonation there is a basic inventory of prosodic portraits of AUs in the various positions within the syntagm: initial, medial and final.

The experimental material for 'cloning' individual prosodic characteristics is either provided by a specially selected comprehensive text recorded by the given speaker or taken from the already available recordings, which are sufficiently varied so as to represent each of the intonational types considered above. The recorded speech corpus is then processed by an expert with the help of a set of standard and original computer techniques of processing speech signals. The expert's task is to provide a prosodic transcription of the sound corpus which includes marking the boundaries between phrases, syntagms and accentual groups, as well as calculating the values of F0, T and A of the different segments of AUs in different positions in relation to the boundaries of the intonation group and the type of the intonation contour. The ultimate goal of an expert's analysis is the creation of a database of personal prosodic portraits of accentual groups, which are then used for synthesizing speech from unrestricted text.

One of the underlying assumptions of the model under discussion was that individual features of prosodic organization manifest themselves both in the structural and the functional aspects. The structural aspect refers to the shape of the tonal configurations coextensive with the separate accentual groups and reveals itself through modifications of the falls and rises of the fundamental frequency (F0) resulting from individual peculiarities in the location of the pitch maxima and minima relative to the duration of the stressed syllable, i.e. the nucleus of the accentual group, and the overall duration of the AU. This is illustrated in Fig. 4 showing F0 contours for 4 different speakers (two males and two females) but for the same intonational type - interrogation, and for the same two words phrase.

The functional aspect is concerned with the distributional characteristics of prosodic units, reflecting the frequency of their occurrence in the speech of the given speaker. Closely related to distributional properties is the feature of prosodic units combinability which

ultimately determines the overall prosodic portrait of a person's speech as a whole and each of the utterances constituting it, in particular.

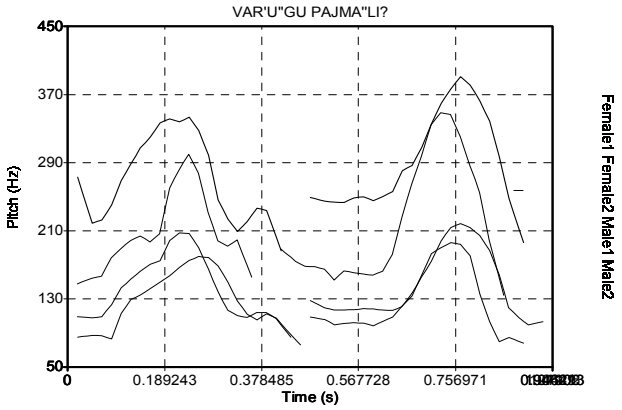

Fig.4. F0 counters for 4 different speakers for the same intonational type of the phrase.

6. System for automation of personal features cloning

The general structure of the *Personal Features Cloning System* is shown in Fig.5. The cloning procedure is based on two types of *texts-corpus* and corresponding them *audio-data* from a speaker: a) for data-driven 'cloning' of individual voice and phometic peculiarities, b) for data-driven 'cloning' of individual features of prosodic organization of the speech. The text's information is an input of *TTS Synthesizer*, while the audio-data is an input of *Speech Signal Parameterization* block. TTS synthesizer is similar to those described above, but its output conveys the information in the form of speech signal parameters labeled according to the allophone sequences. The next block uses the *DTW* algorithm to fulfill a *Labels Transferring* from synthetic speech to natural speech spoken by a certain speaker. The last block carries out the procedure of data-driven *Personal Features Collection*.

7. Conclusion

The analogy drawn here between the biological problem of cloning and the lingua-acoustic problem of synthesizing personified speech from text can be, in our opinion, more than just a metaphor. Firstly, it underlines the general scientific significance and complexity of the task being set. Secondly, it singles out this task into a separate, autonomous class among the other tasks of modern experimental phonetics. And, finally, it stimulates the creation of new specialized technologies, as well as automatic and semi-automatic methods of 'cloning' individual voice in text-to-speech synthesis systems.

The report will be illustrated by the speech samples of the 4 male and 2 female voice-clones in Russian.

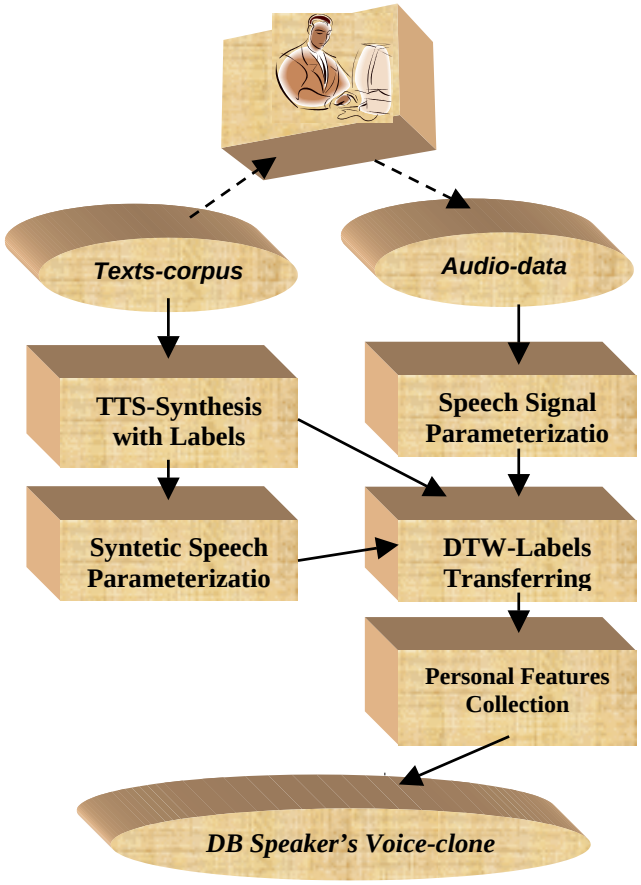

Fig. 5. General structure of personal voice-cloning system

REFERENCES

1. F. CHARPENTIER and E. MOULINES, "Pitch Synhronious Waveform Processing Techniques for TTS Synthesis using Diphones, » in : *Proceedings of Eurospeech'89*, pp. 13-19. Paris,. 1989.

2. S. TAKANO and M. ABE, « A New F0 Modification Algorithm by Maanipulating Harmonics of Magnitude Spectrum, » in : *Proceedings of Eurospeech'99*, pp. 1875-1878. Budapest,. 1999.

3. B. LOBANOV, "The Phonemophon Text-to-Speech System," in: *Proceedings of the XIth ICPhS*, pp. 120-124. Tallinn, 1987.

4. B. LOBANOV and H. KARNEVSKAYA, "MW-Speech Synthesis from Text," in: *Proceedings of the XIIth ICPhS*, pp. 406-409.