Multivoice Text-to-Speech Synthesis for Natural-Language Interfaces of Intelligent Systems

Zahariev V.A.,
Petrovsky A.A.
Belarusian State University
of Informatics and Radioelectronics,
Minsk, Republic of Belarus
Email: zahariev@bsuir.by
Email: palex@bsuir.by

***Abstract*—The paper considers the application of voice conversion technology for developing multivoice text-to-speech synthesis (MVTTS) in natural-language interfaces of intelligent systems. The main features of the proposed integrated architecture of MVTTS system are presented. Voice conversion model based on multiple regression mapping function and Gaussian mixture models, as well as the method of text-independent learning based on hidden Markov models and modified Viterbi algorithm are observed. Experimental evaluation of the effectiveness of the proposed solutions on the characteristics of naturality and similarity of synthesized speech has been done.**

Keywords—text-to-speech system, voice conversion, natural-language interface.

I. Introduction

The natural-language interface is a form of interaction with an intelligent system. This method is optimal for users, because it is the most natural way for communication between humans. At present moment a considerable attention is given to issues connected with natural-language interfaces research and development within the intelligent systems. A text-to-speech synthesis subsystem (TTS), which provides an implementation of a feedback system to the person, takes an important place in the structure of the speech interface system. This type of systems development at the present stage has achieved a significant progress. They have the tasks of not only ensuring intelligibility indicators specified synthesized speech signal, but also requirements on the naturalness of speech, availability of a wide range of prosodic templates, support for multiple languages and different voices announcers. This last aspect is the most important and interesting to study, especially in the context of the transition from the synthesis to voice cloning systems of specific speaker [1]. To solve the task of implementing personalization properties for a speech synthesizer, different speech processing technologies, including voice conversion, are used.

II. Architecture of MVTTS

Voice conversion (VC) is a speech signal processing technology, which allows transforming voice characteristics of the source speaker (SS), contained in speech signal, in characteristics of target speaker (TS) without changing the meaning of the message [2,3]. Voice conversion objects primarily are timbre (spectral envelope) and prosodic (contour of main tone frequency, in another words - pitch) speaker features [4]. To solve the problem of building MVTSS based on VC an approach based on integrated system architecture of TTS and VC was formulated and suggested. We propose to incorporate functional blocks of voice conversion into the TTS composition with at the level of acoustic processor of the system. An information on the voice of the announcer, stored in the DB of acoustic fragments of synthesizer in the parameterized form is used in a function of original. The proposed approach allows to achieve greater cohesion between two types of systems. Multivoice TTS architecture is presented in figure 1.

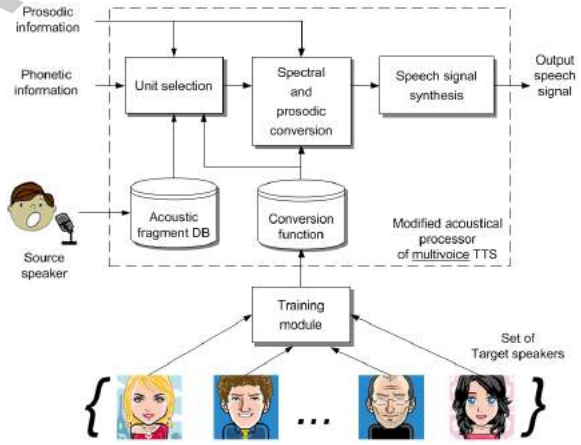

Figure 1. Integrated architecture of MVTTS

Integrated MVTTS architecture allows using available linguistic resources, which the database (DB) contains, acoustic fragments of speech signal kept in the synthesizer (in the parameterized form). A proposed approach permits to achieve greater cohesion between two types of systems. This structure reflects the stages of information processing: aspects of voice conversion are taken into account only while selecting units of compilation; transformation and conversion algorithms of timbral and prosodic information are executed atomically (i.e. signal characteristics are modified only once); compiled and synthesized speech reconstruction also run once immediately after phase conversion characteristics of SS in the TS that reduce the number of alternations of artifacts.

III. Analysis-synthesis speech model

During the review and comparative analysis of the literature on relevant methods of speech signal analysis-synthesis models [5], it was found that the most promising model is a hybrid harmonic plus noise model and its STRAIGHT implementation [6]. It allows decomposing a signal and manipulating independently with three components: main frequency contour (pitch contour), smoothed spectrogram (periodical part), noisy spectrogram (aperiodical part). The analysis of the speech signal is synchronized with the pitch contour and is performed in instantaneous harmonic parameters domain. Then, based on the average estimation between the two instant values of spectral envelope, smoothed time-frequency representation envelope spectrum are taken. Smoothed STRAIGHT-spectrogram is determined according to the expression:

$$P_T(\omega) = \frac{1}{N}\sum_{k=0}^{N-1} |S(\omega, \tau + \frac{kT_0}{N})|^2, \qquad (1)$$

where $P_T(\omega)$ – smoothed stable over time the power spectrum of the signal, provided that the centres localization time Windows are divided into $\frac{T_0}{N}$, N-number of Windows for calculation, $|S(\omega,\tau)|^2$-instant Fourier spectrum for time τ, $\tau + \frac{kT_0}{N}$ – time offset for spectrum analysis of the signal in time $|S(\omega, \tau + \frac{T_0}{2})|^2|$. This representation (1) allows determining accurately the value of periodic and aperiodic signal components best suited for transformation of the spectral envelope. At the later stage it gives the possibility to perform the conversion of each speaker-dependent signal characteristics (envelope spectrum for periodic and aperiodic component and parameter of prosodic in the form of pitch contour) in the simplest way and without additional errors. The method also permits to manipulate the source parameters of excitation, signal recovery of parametric region and provides a synthesis of the signal with low distortion.

IV. Voice conversion based on GMM

The most common paradigm in voice conversion researches now is a statistical model based on multiple Gaussian mixtures models (GMM), which was proposed in the original [7] and its modified versions are presented [8]. Systems based on this model have satisfactory results in terms of the characteristics of the similarities between the converted and the target speech signals.

Training is carried out on the basis of a set of parallel pairs of vectors of SS and TS parameters for which a joint GMM model is built. Expectation vector and covariance matrix available in GMM are used as parameters in the function of conversion, which is represented by the expression:

$$F(\mathbf{x}) = \sum_{q=1}^{Q} p_q(\mathbf{x}_i, \mathbf{y}_{i-1}, \mathbf{x}_{i+1})[\nu_q + \Phi_q \bar{\mathbf{x}}_i^q + \Psi_q \bar{\mathbf{y}}_{i-1}^q + \Omega_q \bar{\mathbf{x}}_{i+1}^q],$$

$$p_q(\mathbf{x}) = \frac{\alpha_q N(\mathbf{x}, \mu_q^x, \Sigma_q^{xx})}{\sum_{j=1}^{M} \alpha_j N(\mathbf{x}, \mu_j^x, \Sigma_j^{xx})}. \qquad (2)$$

where $\mathbf{x}$ – parameters vector of the source speaker, M – number of component mixture, μ_i^x and μ_i^y – expectations vector of the i-th components of the mixture, Σ_i^{xx} – covariance matrix of the source speaker of the i-th components,Σ_i^{yx} – crosscovariance matrix for source and target vectors 1st speaker components, $p_i(\mathbf{x})$ – posteriori probability vector $\mathbf{x}$ i-th component, N – multivariate Gaussian distribution.

Type conversion function (2) is widespread, thanks to a good practical results obtained on its basis [9]. However, it also has some disadvantages. When we try to increase strongly the similarity between speakers, effect of oversmoothing of the voice signal could be observed.

V. Text-independent training method

To build a SS conversion function in TS will a phase of training is required, which includes a stage of mapping vectors of parameters, characterizing the voices of announcers, derived from the analysis and parameterization of the signal. Training can be carried out on the basis of text-independent and text-dependent approaches.

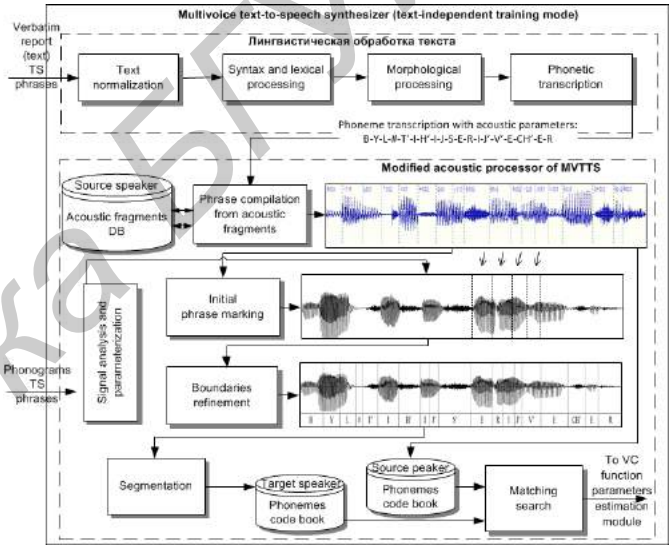

Figure 2. Text-independent training process in MVTTS

While learning, text-dependent SS and TS are required to recite the same text. Common text implies a common phonetic constituents, which vary within certain limits (depending on sex, age, mannerisms, etc. features of the speaker), but essentially the same in terms of phonetics. It gives an opportunity to hold a temporary mapping parameters, vectors and later compare the clusters of acoustic parameters of the subspaces. The downside of this approach is connected with coast of preparation of such parallel corpora.

In this case obvious temporary matching of vectors parameters can't be received by direct contrast. The key advantage of this approach is the convenience and simplicity of customization of the system with this type of training, for the end user. It does not require a certain texts database, you can use any available phonograms or even record the voices in the process of talking on the phone.

The method of text-independent teaching in the context of the MVTSS structure (figure 2.) involves extensive use of linguistic TTS units to undertake the following stages: the text normalization, its syntax, morphological processing, phonemic transcription, as well as further accessed synchronization

markup with phone phonogram target speaker based on data obtained from SS, i.e. speech synthesizer. This fact gives the opportunity to convert subsequently phonemic recognition task units in the flow of speech audio synchronization task and phonetic markup and text blocks of a speech synthesizer. A detailed description of the method could be found in [10].

VI. TEXT-INDEPENDENT TRAINING ALGORITHM

During the alignment stage of training, two concurrent processes are in progress: preparation and analysis of speech and text information on the phonetic and acoustic characteristics of the target speaker. A sequence of phonograms learning phrases $Wav = (w_1, w_2, \dots, w_n)$ and its corresponding recording sequence $Torpho = (t_1, t_2, \dots, t_n)$, are the main information for these processes, where *n* – is the number of phrases of the training set. Next over textual information processor conversion is carried out language spelling of text in an alternative type of L: $Torpho \rightarrow Tphono \in D^{s\times n}$, where *s* – is the number of phonetically distinguishable units in a single phrase, and then the phonetic transformation processor performs phonemic text, F: $Tphono \rightarrow Tallo \in D^{a\times n}$, where *a* – is the number of allophones (combinations of phonemes) in a single phrase, in a sequence of indices of allophones. It should be noted that the alphabet index data matches for all of the speakers, because their composition is strongly typed and is constant for members of a language group. Algorithms that implement the data transformations are described in detail in the literature [11].

Speech signal analysis unit performs over a sequence of phonograms *W* analysis , based on signal model STRAIGHT [4]. To reduce computational complexity of subsequent phases, and the resulting conversion model smoothed spectrum $P_R(w,t)$ for every moment of time is replaced with its envelope in parametric form using linear spectral frequencies (LSP). Thus, the transformation performed by block analysis of speech signal can be defined formally as $\hat{A}$: $Wav \rightarrow \mathrm{Prm} \in D^{p\times m\times n} | \Pr m(m,n) = \{F_0, \Delta F_0, a_1, a_2 ... a_p\}$, where *p* – dimensionality of LSP parameters vector , *m* – number of signal frame, *n* – number of phnogram in training set Next sequence of allophones indexes $Tallo$ and vectors Prm of signal parameters simultaneously are transmitted to the entrances of the block defining the boundaries of allophones, that is establishing an optimal match between them.

This task can be solved efficiently using hidden Markov Model (HMM), in which the sequence $Tallo$ would be a sequence of states, and Prm – the sequence of observations. From the terms of HMM this process consists of linking the best sequence of states from the current sequence of observations for the model. The solution of this problem, formalized in the form of expression, is possible by using iterative Viterbi algorithm.

$$H : (Tallo,\ Prm) \rightarrow (Tallo^{opt},\ Prm),$$

$$\arg\max P(Tallo,\ Prm|\lambda), \lambda \in (A, B, \pi),$$

$$B^{trg} = \{\forall t | Tallo^{opt}(t)\} \Leftrightarrow P^{trg} = \{\forall p | Prm(p)\}$$

where λ – hidden Markov model parameters, A – matrix of states, B – matrix of observations, π– matrix of initial states. Further, by combining statistics all observations on each state, on all phrases training set, it is possible to find matching vectors of parameters relevant to a particular allofonu, thanks to the equality of allophones alphabets in terms of its composition $Ballo^{src}(i) = Ballo^{trg}(i) = Ballo(i) \Rightarrow Prm^{src}(i) \Leftrightarrow Prm^{trg}(i)$. Thus, all of the above steps as a result of the two phases, allow forming joint sequence of vectors of signal parameters on phonetic $Z = (\{Prm^{src}(i),\ Prm^{trg}(i)\}_i), \forall i \in N, i = 1, I$,(where *I* – number of allophones in the database) for its subsequent use in the search model parameters of conversion. More information on the learning stage is described in the works of authors and bibliography [5].

VII. EXPERIMENTAL RESULTS

To evaluate the effectiveness of the proposed methods a series of experiments to characterize the intelligibility and naturalness of synthesized speech by means of method of average opinions (mean opinion score-MOS) was conducted [12].

Experiments were perfomed on phonetically balanced set of phrases, including 90 on audio records of the same proposals for four speakers: two men and two women.In further experiments male speakers are nominally marked as M1 and M2, and female speakers as F1 and F2. The average duration of one phrase was 5-6 seconds. The scaling algorithm was used for temporary alignment based on dynamic programming.

Experimental results are presented in table 1.

Table I. RESULTS OF THE EXPERIMENTAL EVALUATION OF MOS MARKS FOR NATURALITY AND SIMILARITY OF SYNTHESIZED SPEECH

Naturality (MOS)					
	M-	M-F	F-	F-F	Mean
GMM	3,2	2,6	2,9	3,3	3,0
GMM*	3,5	3,2	3,3	3,4	3,4
FW	4,3	3,3	3,7	3,8	3,8
ANN	3,6	3,7	3,7	4,1	3,8
Similarity (MOS)					
	M-	M-F	F-	F-F	Mean
GMM	3,7	3,5	3,6	3,9	3,7
GMM*	3,7	3,6	3,6	3,7	3,7
FW	2,7	2,2	2,8	2,6	2,6
ANN	3,8	3,4	3,8	4,3	3,8

For illustration purposes, the results of the experiment are presented in the form of histograms for characteristics of naturalness (figure 3).

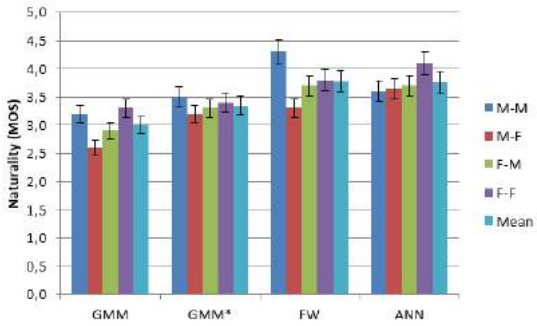

Figure 3. Similarity marks (MOS)

and similarity of synthesis speech (figure 4).

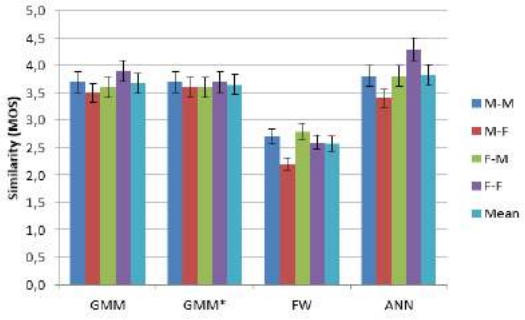

Figure 4. Naturality (MOS)

Experimental results suggest that the proposed method enables to improve the characteristics of naturalness in comparison with the classical method of conversion on the basis of an average of 10 MOS % according to the parameters of naturalness and 5 % according to similarity. Also according to the similarity parameter this method comes short of approach based on spectral weighting and artificial neural networks. This fact can be explained by the fact that proposed method (GMM*) allows to get a stronger, less than the average (more natural) representation of spectral envelope as a result of conversion, while classical statistical techniques (GMM) can significantly average this feature. However, according to degree of similarity, the proposed method exceeds the standard method based on frequency warping (FW) and only is slightly inferior to the method based on artificial neural networks (ANN),the listed methods in the process of achieving great simplicity in training and fewer resources on preparations. High performance of latest justified by a nonlinear mapping displaying features, but require a preliminary design of a neural network and its learning algorithm selection.

VIII. Conclusion

The report deals with aspects of practical implementation of multivoice speech synthesis systems. A new architecture based on speech synthesis systems integration and conversion of voice at the level of modified acoustic processor system synthesis is proposed.

A suggested method and the algorithm have the following advantages: allow to convert input text transcripts for obtaining spelling phonemic text; provide the ability to generate a training set for the original announcer on the basis of the target text of phonemic announcer; all necessary for learning information about SS is kept within the MVTTS as its acoustic resources (database of acoustic standards, dictionaries and rules), or can be obtained by synthesis within the MVTTS on information from TTS.

The proposed methods implemented in the form of software modules could be used as components for the construction of natural-language interfaces of intelligent systems.

References

[1] Лобанов, Б.М. Компьютерный синтез и распознавание речи / Б. М. Лобанов, Л. И. Цирульник – Минск : Белорусская наука, 2008. – 344 с.

[2] Stylianou, Y. Voice transformation: A survey / Y. Stylianou // Proc. of International Conference on Acoustics, Speech and Signal Processing. – Taipei, 2009. – P. 3585–3588.

[3] Shikano, K. Speaker adaptation through vector quantization / K. Shikano, K. Lee, R. Reddy // ICASSP. – 1986. –Vol. 11. –P. 231–237.

[4] Voice conversion: a crtitical survey [Electronic resource] /A.F. Machado, M. Queiroz // Open-acess article. –2010. –Modeof acces: http://www.ime.usp.br/ mqz/SMC2010Voice.–Date of access: 13.05.2013.

[5] Анализаторы речевых и звуковых сигналов / под ред. д.т.н. профессора Петровского А.А. –Минск: Бестпринт, 2009. – 456 с.

[6] Kawahara, H. TANDEM-STRAIGHT: A temporally stable power spectral representation. / H. Kawahara // ICASSP. – 2008. –IEEE. –P. 3933–3936.

[7] Toda, T. Spectral conversion based maximum liklehood estimation considering global variance of converted parameter / T. Toda, A. Black, K. Tokuda. // ICASSP. –2005. –P. 9–12.

[8] Stylinau, Y. Continuous probabilistic transform for voice conversion. / Y. Stylianou // IEEE TSAP. –1998. –№ 6 –P. 131–142.

[9] Erro, D. Weighted frequency warping for voice conversion / D. Erro, A. Moreno // Audio, Speech, and Language Processing, IEEE Transactions on – 2010. –Vol. 18. –P. 543–550.

[10] Захарьев, В.А. Текстонезависимое обучение в системе конверсии голоса на базе скрытых марковских моделей и схемы преобразования буква-фонема / В.А. Захарьев, А.А. Петровский // Цифровая обработка сигналов и её приме-нение (DSPA 2013) —Москва: ИПУ РАН. -–С. 327-332.

[11] Захарьев, В.А. Система синтеза речи по тексту с возможностью настройки на голос целевого диктора / В.А. Захарьев, А.А. Петровский, Б.М. Лобанов // Труды СПИИРАН. – Спб: СПИИРАН. – 2014. – №1(32). – С. 82-98.

[12] Methods for subjective determination of transmission quality // ITU-T Recommendation P. 800: Telephone transmission quality. – 1996. – 37 p.

МУЛЬТИГОЛОСОВОЙ СИНТЕЗ РЕЧИ ПО ТЕКСТУ ДЛЯ ПОСТРОЕНИЯ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ИНТЕРФЕЙСОВ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ

Захарьев В.А., Петровский А.А.

В докладе рассмотрены вопросы применения технологии конверсии голоса для построения мультиголосовых систем синтез речи по тексту (МГСРТ) для создания естественно-языковых интерфейсов интеллектуальных систем. Особенности предлагаемой интегрированной архитектуры МГСРТ на базе технологии конверсии голоса, функция конверсии голоса на основе модели Гауссовых смесей и множественной регрессионной функции отображения, а также метод текстонезависимого обучения на базе скрытых Марковских моделей и модифицированного алгоритма Витерби. Приведены экспериментальные оценки эффективности предлагаемых решений по характеристикам узнаваемости и натуральности синтезируемой речи