First experiments on text-to-speech system personification

Zdeněk Hanzlíček, Jindřich Matoušek, and Daniel Tihelka

University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
{ zhanzlic, jmatouse, dtihelka } @ kky.zcu.cz

Abstract. In the present paper, several experiments on text-to-speech system personification are described. The personification enables TTS system to produce new voices by employing voice conversion methods. The baseline speech synthetizer is a concatenative corpus-based TTS system which utilizes the unit selection method. The voice identity change is performed by the transformation of spectral envelope, spectral detail and pitch. Two different personification approaches are compared in this paper. The former is based on the transformation of the original speech corpus, the latter transforms the output of the synthesizer. Specific advantages and disadvantages of both approaches are discussed and their performance is compared in listening tests.

Key words: TTS system personification; speech synthesis; voice conversion

1 Introduction

Within the concatenative corpus-based speech synthesis framework, a new voice can be obtained by recording a new large speech corpus by the demanded speaker. From that corpus, containing several thousands of utterances, a new unit inventory is created and used within the synthesis process [1]. However, recording of such a great amount of speech data is a difficult task. Usually, a professional speaker is required.

Alternatively, text-to-speech system personification [2] enables this system to produce new voices by employing voice conversion methods. Much fewer speech data are necessary. Our voice conversion system [3] converts spectral envelope and pitch by probabilistic transformation functions; moreover, spectral detail is transformed by employing residual prediction method.

Two different personification approaches are described and compared in this paper. The former is based on the original speech corpus transformation, the latter transforms the output of the synthesizer. Specific advantages and disadvantages of both approaches are discussed and the performance is compared by using preference listening tests.

The paper is organised as follows. In Section 2, the baseline TTS system planned to be personified is described. In Section 3, the voice conversion methods are specified. Section 4 deals with the TTS system personification task. Section 5 describes our first personification experiments. In Section 6, the results are discussed and future work is outlined.

2 Baseline TTS system

The text-to-speech system ARTIC employed in our personification experiments was in detail described in [1]. It has been built on the principles of concatenative speech synthesis. Primarily, it consists of three main modules: acoustic unit inventory, text processing module and speech production module. It is a corpus-based system, i.e. large and carefully prepared speech corpora are used as the ground for the automatic definition of speech synthesis units and the determination of their boundaries as well as for unit selection technique.

Our TTS system was designed for the Czech language, nevertheless many of its parts are language-independent. For our personification experiments, a female speech corpus containing 5,000 sentences (about 13 hours of speech) was employed. The block diagram of our TTS system is shown in Fig. 1.

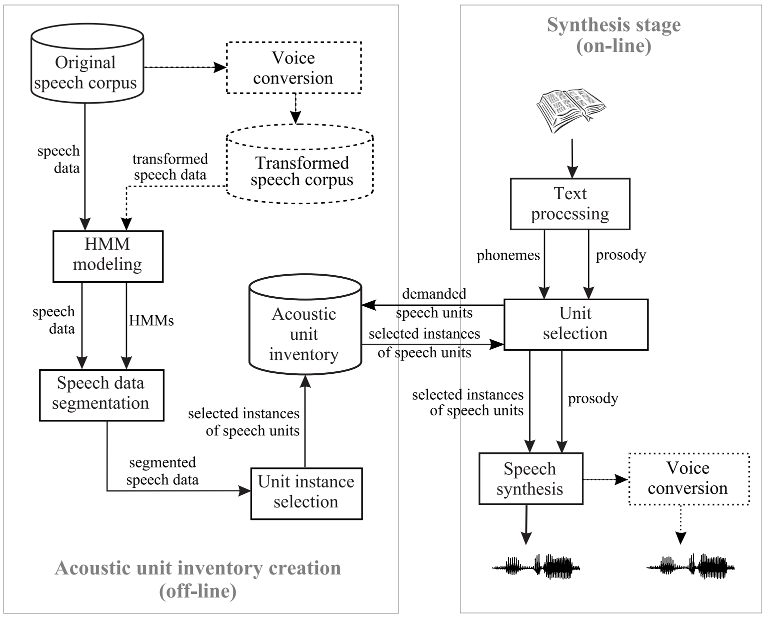

Fig. 1. A scheme of our TTS-system ARTIC including the both personification approaches – see dashed and dotted blocks.

3 Voice conversion system

The voice conversion system utilized for the aforementioned system personification was introduced in [3]. A simplified version of that system is described in this section. For the training of transformation functions, parallel utterances (i.e. pairs of source and target speakers' utterances) are employed. Voiced speech is analysed pitch synchronously;

each segment is three pitch periods long and the shift of analysis window is one pitch period. Unvoiced segments are 10 msec long with 5 msec overlap. The spectral envelope of each frame is obtained by using the true envelope estimator [4] and represented by its line spectral frequencies (LSFs). The parameter order is selected individually for each speaker in order that the average envelope approximation error is lower than predefined threshold. Moreover, spectral detail is obtained as a complement of the spectral envelope into the full spectrum. In case of linear prediction analysis, the spectral detail corresponds to the residual signal spectrum. The LSF parameters and the fundamental frequency are transformed by probabilistic transformation functions. The spectral detail is estimated by a residual prediction method.

3.1 Parameter transformation

Nowadays, the probabilistic (GMM-based) transformation [5] is the most often used transformation function in VC systems. The interrelation between the time-aligned source and target speaker's LSFs (x and y, respectively) is described by a joint Gaussian mixture model with M mixtures Ω_m^{p}

$$p(x,y) = \sum_{m=1}^{M} p(\Omega_m^{\mathrm{p}}) \mathcal{N}\left\{ \begin{bmatrix} x \\ y \end{bmatrix} ; \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix}, \begin{bmatrix} \Sigma_m^{(x)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(y)} \end{bmatrix} \right\}. \tag{1}$$

All unknown parameters are estimated by employing the expectation-maximization algorithm. The transformation function is defined as conditional expectation of target y given source x

$$\tilde{y} = \sum_{m=1}^{M} p(\Omega_m^{\mathrm{p}} \mid x) \Big[\mu_m^{(y)} + \Sigma_m^{(yx)} \big(\Sigma_m^{(x)}\big)^{-1} (x - \mu_m^{(x)}) \Big], \tag{2}$$

where $p(\Omega_m^{\mathrm{p}} \mid x)$ is the conditional probability of mixture Ω_m^{p} given source parameter vector x

$$p(\Omega_m^{\mathrm{p}} \mid x) = \frac{p(\Omega_m^{\mathrm{p}}) \mathcal{N}\{x; \mu_m^{(x)}, \Sigma_m^{(x)}\}}{\sum_{i=1}^{M} p(\Omega_i^{\mathrm{p}}) \mathcal{N}\{x; \mu_i^{(x)}, \Sigma_i^{(x)}\}}. \tag{3}$$

3.2 F$_0$ transformation

Analogically to the case of parameter conversion, time-aligned source and target instantaneous f$_0$ values $f^{(x)}$ and $f^{(y)}$ are described with a joint GMM

$$p\left(f^{(x)}, f^{(y)}\right) = \sum_{s=1}^{S} p(\Omega_s^{\mathrm{f}}) \mathcal{N}\left\{ \begin{bmatrix} f^{(x)} \\ f^{(y)} \end{bmatrix} ; \begin{bmatrix} \mu_s^{(fx)} \\ \mu_s^{(fy)} \end{bmatrix}, \begin{bmatrix} \sigma_s^{(fx)} & \sigma_s^{(fxfy)} \\ \sigma_s^{(fyfx)} & \sigma_s^{(fy)} \end{bmatrix} \right\}. \tag{4}$$

Again, the converted fundamental frequency $\tilde{f}^{(y)}$ is given as the conditional expectation of target $f^{(y)}$ given source $f^{(x)}$

$$\tilde{f}^{(y)} = \sum_{s=1}^{S} p\big(\Omega_s^{\mathrm{f}} \mid f^{(x)}\big) \Big[\mu_s^{(fy)} + \frac{\sigma_s^{(fyfx)}}{\sigma_s^{(fx)}} \big(f^{(x)} - \mu_s^{(fx)}\big) \Big]. \tag{5}$$

3.3 Spectral details transformation

Spectral detail is also very important for speaker identity perception. It is a complement of the spectral envelope into the full spectrum and consists of amplitude and phase parts – $A(\omega)$ and $\varphi(\omega)$, which are converted separately. Its transformation usually utilizes the relationship to the shape of spectral envelope, e.g. by employing codebooks [6].

The training stage starts with the clustering of training parameter vectors y into Q classes $\mathit{\Omega}^{\mathrm{r}}_{q}$; k-means algorithm is employed. Each class $\mathit{\Omega}^{\mathrm{r}}_{q}$ is represented by its centroid $\bar{y}_q$ and covariance matrix S_q. The pertinence of parameter vector y_n to class $\mathit{\Omega}^{\mathrm{r}}_{q}$ is defined as

$$w(\mathit{\Omega}^{\mathrm{r}}_{q} \mid y_n) = \frac{\big[(y_n - \bar{y}_q)^{\mathrm{T}} S_q^{-1} (y_n - \bar{y}_q)\big]^{-1}}{\sum_{i=1}^{Q} \big[(y_n - \bar{y}_i)^{\mathrm{T}} S_i^{-1} (y_n - \bar{y}_i)\big]^{-1}}. \tag{6}$$

All target speaker's training data are uniquely classified into those classes. For each class $\mathit{\Omega}^{\mathrm{r}}_{q}$, a set $\mathbb{R}_q$ of pertaining data indexes is established

$$\mathbb{R}_q = \Big\{k;\ q = \operatorname*{arg\,max}_{q=1\ldots Q} w(\mathit{\Omega}^{\mathrm{r}}_{q} \mid y_k)\Big\}. \tag{7}$$

Within each parameter class $\mathit{\Omega}^{\mathrm{r}}_{q}$, the training data are divided into L_q subclasses $\mathit{\Omega}^{\mathrm{r}}_{q,\ell}$ according to the instantaneous fundamental frequency f_0. Each subclass $\mathit{\Omega}^{\mathrm{r}}_{q,\ell}$ is described by its centroid $\bar{f}^{(y)}_{q,\ell}$. The data belonging into this subclass are defined using a set $\mathbb{R}_{q,\ell}$ of corresponding data indices

$$\mathbb{R}_{q,\ell} = \Big\{k;\ k \in \mathbb{R}_q\ \wedge\ \ell = \operatorname*{arg\,min}_{\ell=1\ldots L_q} \big|f^{(y)}_k - \bar{f}_{q,\ell}\big|\Big\}. \tag{8}$$

For each subclass $\mathit{\Omega}^{\mathrm{r}}_{q,\ell}$, a typical spectral detail is determined as follows. Typical amplitude spectrum $\hat{A}^{(y)}_{q,\ell}(\omega)$ is determined as the weighted average over all amplitude spectra $A^{(y)}_n(\omega)$ belonging into that subclass

$$\hat{A}^{(y)}_{q,\ell}(\omega) = \frac{\sum_{n \in \mathbb{R}_{q,\ell}} A^{(y)}_n(\omega) w(\mathit{\Omega}^{\mathrm{r}}_{q} \mid y_n)}{\sum_{n \in \mathbb{R}_{q,\ell}} w(\mathit{\Omega}^{\mathrm{r}}_{q} \mid y_n)} \tag{9}$$

and the typical phase spectrum $\hat{\varphi}^{(y)}_{q,\ell}(\omega)$ is selected

$$\hat{\varphi}^{(y)}_{q,\ell}(\omega) = \varphi^{(y)}_{n^*}(\omega) \qquad n^* = \operatorname*{arg\,max}_{n \in \mathbb{R}_{q,\ell}} w(\mathit{\Omega}^{\mathrm{r}}_{q} \mid y_n). \tag{10}$$

During the transformation stage, for the transformed parameter vector $\tilde{y}_n$ and fundamental frequency $\tilde{f}^{(y)}_n$, the amplitude spectrum $\tilde{A}^{(y)}_n(\omega)$ is calculated as the weighted average over all classes $\mathit{\Omega}^{\mathrm{r}}_{q}$. However, for each class $\mathit{\Omega}^{\mathrm{r}}_{q}$, only one subclass $\mathit{\Omega}^{\mathrm{r}}_{q,\ell}$ is selected in such a way that its centroid $\bar{f}^{y}_{q,\ell}$ is the nearest to frequency $\tilde{f}^{(y)}_n$

$$\tilde{A}^{(y)}_n(\omega) = \sum_{q=1}^{Q} w(\mathit{\Omega}^{\mathrm{r}}_{q} \mid \tilde{y}_n) \hat{A}^{(y)}_{q,\ell_q}(\omega) \qquad \ell_q = \operatorname*{arg\,min}_{\ell=1\ldots L_q} \big|\tilde{f}^{(y)}_n - \bar{f}^{(y)}_{q,\ell}\big|. \tag{11}$$

The phase spectrum $\tilde{\varphi}_n(\omega)$ is selected from the parameter class $\Omega_q^{\mathrm{r}*}$ with the highest weight $w(\Omega_q^{\mathrm{r}} \mid \tilde{y}_n)$ from that subclass $\Omega_{q,\ell}^{\mathrm{r}*}$ having the nearest central frequency $\bar{f}_{q,\ell}^{(y)}$

$$\tilde{\varphi}_n^{(y)}(\omega) = \hat{\varphi}_{q^*,\ell^*}^{(y)}(\omega) \qquad \begin{aligned} q^* &= \operatorname*{arg\,max}_{q=1\ldots Q} w(\Omega_q^{\mathrm{r}} \mid \tilde{y}_n) \\ \ell^* &= \operatorname*{arg\,min}_{\ell=1\ldots L_{q^*}} \left|\tilde{f}_n^{(y)} - \bar{f}_{q^*,\ell}^{(y)}\right|. \end{aligned} \tag{12}$$

4 TTS system personification

4.1 Personification approaches

In principle, two main approaches to concatenative TTS system personification exists.

1. *Transformation of the original speech corpus* – a new unit inventory is created from the transformed corpus. Thus for each new voice an individual unit inventory is created and ordinarily used for the speech synthesis.
2. *Transformation of TTS system output* – a transformation module is added to the TTS system. The generation of the new voice is performed in two stages: synthesis of the original voice and transformation to the target voice.

Each of these approaches has specific advantages and also disadvantages. The approach based on original corpus transformation can be characterized as follows:

+ The converted corpus can be checked and poorly transformed utterances rejected. Thus the influence of conversion failure can be suppressed.
+ The synthesis process is straightforward and it is not delayed by additional transformation computation.
– The preparation of new voice is time consuming – the whole corpus has to be converted and a new unit inventory built.
– Huge memory requirements for storing several acoustic unit inventories, especially in cases when more different voices should be alternatively synthesized.

Properties of the second approach can be briefly summarized:

+ A new voice can be simply and quickly acquired, only a new set of conversion functions has to be added.
+ Lower memory requirements – only the original unit inventory and conversion functions for other voices have to be stored.
– The resulting system works slower – an extra computation time for transformation is needed.

4.2 Data origin

In our conversion system, parallel speech data is necessary for the training of conversion function. Within the TTS system personification framework, source speaker's speech data can be obtained in two different ways

- *Natural source speech data* – Source speaker's utterances are selected from the original corpus. The recording of additional utterances by the source speaker is less suitable, especially in cases when a long time has elapsed since original corpus recording, because his/her voice could change since that time.
- *Synthetised source speech data* – Source's speaker utterances are generated by the TTS system. This is necessary in cases when target speaker's utterances are given, but not involved in the source corpus. Moreover, none of both speakers is available for an additional recording.

Considering the training and transformation stage consistency, a natural training data seems to be preferable for the source corpus conversion. However, in the case of TTS system output transformation, a sythetised source training data is more suitable.

5 Experiments

The performance of a conversion system can be evaluated by using so-called performance indices

$$P_{\mathrm{par}} = 1 - \frac{\sum_{i=1}^{N} \mathcal{D}(\tilde{y}_n, y_n)}{\sum_{i=1}^{N} \mathcal{D}(x_n, y_n)} \qquad P_{\mathrm{sp}} = 1 - \frac{\sum_{i=1}^{N} \mathcal{D}\big(\widetilde{A}_n^{(y)}(\omega), A_n^{(y)}(\omega)\big)}{\sum_{i=1}^{N} \mathcal{D}\big(A_n^{(x)}(\omega), A_n^{(y)}(\omega)\big)}, \qquad (13)$$

where x_n, y_n and $\tilde{y}_n$ are source, target and transformed parameter vectors, $A_n^{(x)}(\omega)$, $A_n^{(y)}(\omega)$ and $\widetilde{A}_n^{(y)}(\omega)$ are corresponding spectral envelopes and $\mathcal{D}$ is usually the Euclidean distance.

The higher are the values of parameter and spectral performance index, the higher is the similarity between transformed and target utterances in comparison with the original similarity between source and target utterances.

In addition to those objective mathematical rates, the speech produced by conversion system or by the personified TTS system can be evaluated in listening tests. For comparison of several system setups a preference test can be employed.

In our experiments, two nonprofessional target speakers recorded 50 quite short sentences (about 6–8 words long), which were selected from the corpus mentioned in Section 2. Thus, parallel training data was available. Within the training stage, 40 utterance pairs were used for the estimation of conversion function parameters.

5.1 The influence of data origin

Regardless of the personification approach, source speaker's training data can either be natural or synthetised by the TTS system (or both together, but that case was not taken into account). Hereinafter, we use notation NTD/STD function for conversion functions trained by using natural/synthetised source training data.

A question arises whether the conversion function trained on natural data could be used for synthetised speech transformation and vice versa. Thus, NTD and STD conversion functions were trained and employed for the transformation of both natural and synthetised speech. An objective comparison of NTD and STD performance, based

Table 1. The influence of training data origin: natural or synthetised.

Training data	Testing data	Male 1		Male 2	
		P_{par}	P_{sp}	P_{par}	P_{sp}
natural	natural	0.239	0.230	0.354	0.344
synth.	synth.	0.242	0.233	0.331	0.324
synth.	natural	0.194	0.194	0.357	0.347
natural	synth.	0.209	0.200	0.325	0.316

on performance indices, is presented in Table 1. The utterances, that were not included in the training set, were used for this assessment.

Moreover, informal listening test was carried out. 10 participants listened to the pairs of utterances transformed by NTD and STD function. The natural and synthetised utterances from both target speakers were evenly occured in the test. In each testing pair, the listeners should select a preferred utterance according to the overall voice quality. The similarity to the real target speaker's voice was not taken into account. The results of this test are presented in Figures 2 and 3.

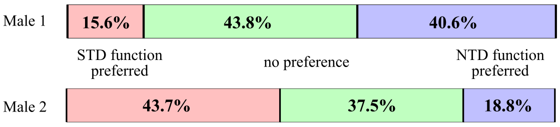

Fig. 2. Preference listening test: Synthetised speech transformation.

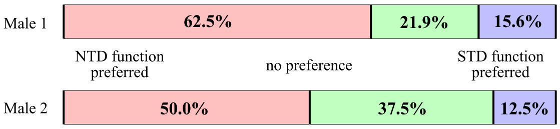

Fig. 3. Preference listening test: Natural speech transformation.

The results of the mathematical evaluation and the listening test are consistent. For both speakers, natural speech is better transformed by NTD function. The results for synthetised speech transformation differ for particular speakers. However, the differences between the utterances were mostly insignificant.

5.2 Personification approaches comparison

For the comparison of described personification approaches another preference listening test was employed. Again, participants listened to the pairs of utterances produced by the TTS systems personified by the source corpus transformation (approach 1) and synthetiser output transformation (approach 2). The results are presented on Figure 4. For both speakers approach 2 was preferred.

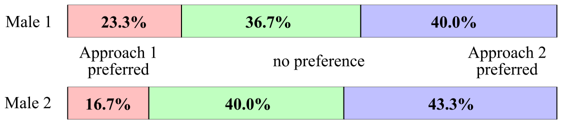

Fig. 4. Preference listening test: Personification approaches comparison.

6 Conclusion

In this paper, two different approaches to the TTS system personification were compared. The former is based on the original speech corpus transformation and a new unit inventory creation, the latter transforms the output of the original TTS system. In listening tests the corpus transformation approach revealed to be slightly preferred. However, the differences were not too significant. Thus, both approaches are well applicable. Their specific advantages and disadvantages should be considered for concrete applications.

7 Acknowledegment

Support for this work was provided by the Ministry of Education of the Czech Republic, project No. 2C06020, and by the Grant Agency of the Czech Republic, project No. GAČR 102/09/0989.

References

1. Matoušek, J., Tihelka, D. and Romportl, J.: Current State of Czech Text-to-Speech System ARTIC. Proceedings of TSD, LNAI 4188. Springer, Berlin (2006) 439–446.
2. Kain, A. and Macon, M. W.: Personalizing a Speech Synthesizer by Voice Adaptation. Proceedings of SSW. Blue Mountains, Australia (1998) 225–230.
3. Hanzlíček, Z. and Matoušek, J.: Voice Conversion based on Probabilistic Parameter Transformation and Extended Inter-Speaker Residual Prediction. Proceedings of TSD, LNAI 4629. Springer, Berlin (2007) 480–487.
4. Villavicencio, F., Röbel, A. and Rodet X.: Improving LPC Spectral Envelope Extraction of Voiced Speech by True-Envelope Estimation. Proceedings of ICASSP. Toulouse, France (2006) 869–872.
5. Stylianou, Y., Cappé, O. and Moulines, E.: Continuous Probabilistic Transform for Voice Conversion. IEEE Trans. on Speech and Audio Processing, Vol.6, No.2 (1998) 131–142.
6. Kain, A.: High Resolution Voice Transformation. Ph.D. thesis, Oregon Health & Science University, Portland, USA (2001).