A Comparison of Voice Conversion Methods for Transforming Voice Quality in Emotional Speech Synthesis

Oytun Türk, *Marc Schröder*

DFKI GmbH, Language Technology Lab, Saarbrücken and Berlin, Germany

oytun.tuerk@dfki.de, marc.schroeder@dfki.de

Abstract

This paper presents a comparison of methods for transforming voice quality in neutral synthetic speech to match cheerful, aggressive, and depressed expressive styles. Neutral speech is generated using the unit selection system in the MARY TTS platform and a large neutral database in German. The output is modified using voice conversion techniques to match the target expressive styles, the focus being on spectral envelope conversion for transforming the overall voice quality. Various improvements over the state-of-the-art weighted codebook mapping and GMM based voice conversion frameworks are employed resulting in three algorithms. Objective evaluation results show that all three methods result in comparable reduction in objective distance to target expressive TTS outputs whereas weighted frame mapping and GMM based transformations were perceived slightly better than the weighted codebook mapping outputs in generating the target expressive style in a listening test.

Index Terms: voice quality transformation, voice conversion, emotional speech synthesis

1. Introduction

Development of techniques for generating synthetic speech using a large number of speech recordings resulted in high quality systems available in both research and commercial domains. The functionality to add and control expressiveness in synthetic speech has become an important research aspect in such systems to further improve naturalness in human-machine interaction. The conventional approach to generate expressive synthetic speech in unit selection systems employed collection of a separate large corpora in the target expressive style which is a time consuming and tedious task. An alternative approach is to employ voice modification and conversion techniques to adapt neutral synthesis output to match the target styles. The adaptation can be performed using a considerably smaller number of recordings in the target expressive style, enabling inclusion of new expressive styles in the system without much additional effort.

The efforts towards voice modification and conversion to generate a target expressive style in TTS focus on two major problems. The first problem is the transformation of voice quality which is known to convey significant variation across different styles. However, explicit modeling and modification of voice quality parameters is still an open research question and needs future improvements to provide a fully automatic parametric framework. Alternatively, voice conversion techniques can be used to transform the overall spectral characteristics for realizing corresponding voice quality changes implicitly in the spectral conversion function. The second problem involves generation and realization of the appropriate prosody patterns that fit the target expressive style. This paper focuses on the first problem to provide a voice quality transformation framework using conventional speaker identity conversion methods.

Voice conversion researchers employ different machine learning techniques to automatically map a given source speaker´s spectral characteristics to that of a target speaker. Codebook mapping [1] and GMM based [2, 3] approaches have become two extensively used methods. Codebook mapping models the spectral envelope transformation function by clustering the source and target training data and by modifying the source spectral envelope by a filter estimated from the source and the corresponding target cluster center. The major drawback of this approach is known to be the discontinuity problems due to local modeling of the transformation function. Weighted codebook mapping resulted in significant performance improvements [4, 5] and provided the framework for a number of commercial applications including cross-language dubbing and singing voice transformation. A further improvement includes frame selection/mapping approaches to generate the target characteristics in more detail [6].

GMM based voice conversion algorithms provide a parametric framework to model the relationship between the source and target acoustic spaces. The spectral transformation function can be estimated using a GMM trained on either source spectral feature vectors [2] or source-target spectral feature vectors jointly [3]. The major drawbacks of GMM based approaches include over-smoothing problems [7] and discontinuities in the transformation filter across consecutive speech frames. However, it is possible to reduce these effects by combining codebook mapping and GMM based methods [8] and by employing temporal smoothing of the transformation filter [9].

The selection of appropriate signal processing algorithms to manipulate voice quality and prosody characteristics is crucial since additional processing distortion is a strictly limiting factor in TTS applications. In this paper, we use LP inverse filtering followed by FD-PSOLA [10] and frequency domain filtering to perform voice quality and prosody transformations. Specifically, we compare the performances of three methods for voice quality transformation in generating emotional synthetic speech using the MARY TTS platform [11]. The first method employs improved weighted codebook mapping [5] where the weighted codebook mapping procedure in [4] has been extended to automatically detect and eliminate outliers in the training data. The outlier elimination procedure makes use of the distributions of various distance measures between the source and target acoustic features. The second voice quality transformation method uses a direct frame mapping approach in which a more detailed vocal tract transformation function is estimated using individual speech frames and context information [6]. The third method is based on classical joint source-target GMM [3]. Both the weighted frame mapping and GMM based methods were extended with the outlier elimination procedure. We have also applied temporal smoothing of the transformation function similar to [9] to reduce possible discontinuities in the transformation filter as required. Since the focus of the recent work is not on prosody transformation, we simply generate the target prosody by CARTs trained on the expressive data to provide target

pitch and duration values. The target values are super-imposed during voice conversion to modify pitch and timing. No models of intensity are employed although it may help in improving performance further in various expressive styles.

The outline of the paper is as follows: Section 2 reviews the MARY expressive TTS system which is used in generating neutral speech as input to the voice conversion algorithm and delivering expressive speech samples as a ground reference for performance evaluations. The three voice quality transformation algorithms are described next along with a brief summary of expressive prosody generation. Section 3 presents objective and subjective evaluation procedures and results. Finally, the paper is concluded with a discussion of findings and future research directions in Section 4.

2. Method

2.1. MARY expressive TTS system

We use the unit selection system of the MARY TTS platform [11] for our experiments. It implements a standard unit selection algorithm, generating speech by concatenating *units* – small snippets of audio recordings – selected from natural recordings of a given speaker and style. First, a *target* is predicted from the text, consisting usually of linguistic descriptors such as the phone chain, part-of-speech information etc., as well as a symbolic or parametric description of target prosody. Potentially, this *target* could also include information about the intended emotion or speaking style. From the available unit inventory of a given voice, units are selected to match the target as closely as possible; at the same time, they must also fit to the neighboring units. Dynamic programming is used to find an optimal path through the candidate units to jointly minimize the target costs and the join costs. The MARY system uses diphone-sized units, with an option to fall back to half-phones if no suitable diphones are available.

2.2. Database

In the standard unit selection framework, expressive synthetic speech is simply generated by using recordings of expressive speech as the voice database. We have recorded several voice databases with the same speaker, a professional German actor, who is speaking in a different style for each voice database.

A standard, "neutral" database was recorded as a baseline, consisting of 3000 phonetically and prosodically balanced sentences selected from the German Wikipedia [12]. The same speaker produced several smaller, expressive databases, notably a *cheerful*, an *aggressive*, and a *depressed* voice. For each voice, 400 phonetically balanced Wikipedia sentences and approximately 200 style-specific sentences were recorded, based on the concept of domain-oriented synthesis [13]: within a certain "domain" (such as, here, typical "cheerful" sentences), the voice will have high quality; outside the domain, it will at least be intelligible.

Recordings were done with 24 bit at 44.1 KHz, directly into a laptop computer, using the tool Redstart from the MARY system and down-sampled to 16 KHz during voice generation. The voices used in this paper were built fully automatically, using the MARY voice import toolkit [14]; manual correction of the labels is currently under way, and will result in voices with more reliable quality.

In training spectral envelope transformation functions, a 200-sentence subset of the parallel recordings was used for each expressive style. The transformation functions were trained using the neutral recordings as the source and the expressive recordings as the target. For testing, three utterances were synthesized using the neutral voice along with the realized phoneme boundaries and pitch contour. These three utterances were also synthesized using the expressive voice as well for generating the target prosody pattern using the corresponding CART trained with the expressive databases. The f0 contour and the durations as predicted by the expressive CART and the phoneme boundary information for the expressive outputs were used as input to the voice conversion algorithm to determine time varying pitch and time scaling factors to be realized during FD-PSOLA based prosody modifications.

2.3. Voice quality transformation

2.3.1. Weighted codebook mapping

The first spectral envelope transformation algorithm is based on the weighted codebook mapping algorithm STASC [4] including the refinements proposed in [5]. The algorithm extracts average source and target LSF vectors for each source and target phoneme pair in a parallel training database and saves them in a codebook file. In the transformation stage, the source LSF vectors for each input speech frame are matched with source codebook entries using an LSF based distance measure. The converted LSF vector is estimated as a weighted average of the corresponding target codebook entries using the inverse of the source distance values as a weighting factor. This procedure is sensitive to alignment accuracy similar to all parallel voice conversion training algorithms. In order to eliminate codebook entries that correspond to misaligned labels automatically, we use the outlier elimination method described in [5]. The outlier elimination computes four confidence measures based on the distributions of LSF, f0, root-mean-squared energy, and duration differences between the source and the target codebook entries. A single Gaussian is fitted to each distribution and the entries that have larger difference from the Gaussian mean by an automatically determined threshold are eliminated. This helps to exclude relatively different source and target pairs from the codebook, resulting in a more stable and less discontinuous transformation function. In the transformation stage, source codebook entries can be used to estimate the input vocal tract spectrum from the source codebook as described in [4]. Then, the input spectrum is inverse filtered with the source vocal tract spectrum estimate rather than the direct LP estimate to improve smoothness of the transformation filter across consecutive speech frames.

2.3.2. Weighted frame mapping

An extension to the weighted codebook mapping algorithm described in the previous sub-section is proposed in the first author´s previous work for speaker identity conversion [6]. It includes mapping of LSF vectors extracted from source and target speech frames directly rather than averaging the source and target LSF vectors for each observed phoneme. This results in a more detailed modeling of the mapping between the source and the target vocal tract characteristics. As a consequence of generating codebook entries from speech frames, the codebook sizes become significantly larger - typically 10 to 20 times of the phoneme-averaged codebook size. In the transformation stage, the codebook search procedure is constrained with context information in order to reduce computational requirements and to improve robustness in frame based codebook matching. Except the employment of context information and searching in a large, speech frame

based codebook, the frame mapping based algorithm consists of identical steps as in weighted codebook mapping.

A two stage smoothing procedure is then employed since the frame mapping procedure is more likely to result in discontinuities due to the estimation of vocal tract transformation filter directly from speech frame LSFs. The first smoothing stage involves inverse filtering with source vocal tract filter estimated from the source codebook as in weighted codebook mapping. In addition, we apply temporal smoothing of the transformation filter similar to [9] to further reduce discontinuities in the filter in consecutive speech frames. Temporal smoothing was performed off-line by weighted averaging of filter spectrum bins using a Gaussian window in an additional pass. A total of three (previous, current, and next frame) filters are used in the smoothing process.

2.3.3. Joint source-target GMM

The GMM framework [2, 3] has been widely used for speaker identity conversion in speech and expressive style transformation in TTS [15]. In our implementation, we fit a GMM with 40 mixture components to the joint source and target LSF acoustic space as described in [3] using the expectation-maximization (EM) algorithm. The minimum number of EM iterations was set to 100 and the algorithm automatically decided when to quit iterations by considering the average change in mixture means in consecutive iterations.

2.4. Prosody transformation

In the MARY unit selection system, prosody models are trained for each voice using the speech corpus. Regression trees are trained for phone durations and for three f0 values per syllable: initial, medial, and final f0 value in the voiced section of the syllable. In order to select suitable questions at the decision nodes, a large range of linguistic features are predicted using the MARY system for each phone in each sentence in the speech database. The tool *wagon* from the Edinburgh speech tools is then used to compute the trees. While this method is very simple, it can be expected to capture the typical prosody of a given voice. In standard unit selection, the acoustic models trained in this way are used to predict the target prosody for the computation of *acoustic target costs* in the unit selection process.

During transformation, we use the FESTIVAL_UTT output as generated by MARY for providing target f0 and duration values as well as corresponding phonetic labels for expressive voices. The expressive labels are aligned with the corresponding neutral TTS output labels using dynamic programming since there can be minor differences in these two due to silence insertions and deletions. Time varying duration and pitch scaling factors are estimated for each pitch synchronous frame of the neutral TTS output provided that it corresponds to a voiced frame. Otherwise, both pitch and duration scaling factors are set to unity. The time varying pitch and duration scaling factors are used in the transformation stage in FD-PSOLA to generate the CART based expressive pitch and duration estimates at the output.

2.5. Outline of voice quality transformation system for expressive TTS

The flowchart of the proposed system for voice quality transformation is shown in Figure 1. Using the MARY Voice Import Toolkit [14], a neutral unit selection voice is created using the large database described in Section 2.2. Expressive prosody predicting CARTs are trained using the small expressive databases. Voice conversion training is then performed using 200 parallel utterances from the neutral database and the expressive databases with the three methods described in Section 2.3. In the transformation stage, neutral TTS output is generated from text using MARY TTS. The vocal tract spectrum of the neutral TTS output is transformed using the voice conversion models in the frequency domain. The target f0 and duration values as predicted by the expressive CARTs are provided as input to the FD-PSOLA based prosody transformation algorithm. We followed a two-step approach to obtain voice quality and prosody transformed output based on our informal observations that the overall quality was slightly better as compared to a single-step approach: Vocal tract transformation is performed in the first step followed by FD-PSOLA based prosody modifications in the second step.

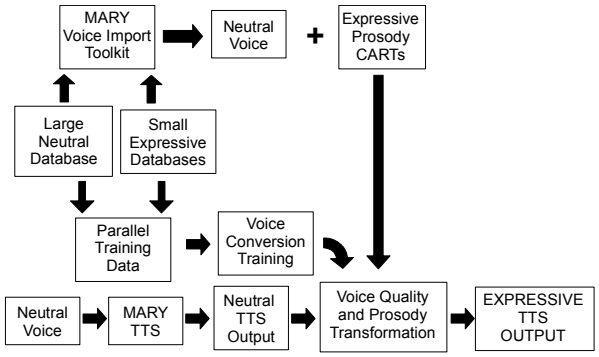

Figure 1: *Flowchart of the voice quality transformation system for expressive TTS.*

3. Evaluations

3.1. Objective evaluation

We have used the root-mean-squared error (RMSE) of Bark-scaled LSF values as estimated from TTS outputs or transformed TTS outputs. RMSE between the transformed TTS outputs and the corresponding TTS outputs generated by the expressive synthesizer is estimated using:

$$RMSE_i = \sqrt{\frac{1}{P}\sum_{k=0}^{P-1}(e_i(k) - t_{m(i)}(k))^2} \qquad (1)$$

where P is the linear prediction order, e and t are the mapped Bark-scaled LSF vectors of expressive TTS outputs and transformed TTS outputs respectively, i is the speech frame index and $m(i)$ is the mapping of speech frame indices using phonetic alignment information. An LP order of 18 was used for 16 KHz recordings. The mean RMSE values shown in Table 1 are computed excluding the silence periods in the beginning and at the end of the signals. Comparing the original distances between neutral and expressive TTS outputs on the last row of the table, we observe that all three algorithms result in a reduction of the RMSE based measure. Weighted frame mapping results in slightly better performance indicating that modelling the mapping between the source and the target in more detail leads to additional reduction in neutral to expressive distance values. The objective performances of weighted codebook mapping and GMM are similar.

Table 1: *RMSE of Bark-scaled LSF values.*

Method	Aggressive	Cheerful	Depressed	All
Codebook	0.54	0.53	0.50	0.52
Frame	0.52	0.51	0.46	0.50
GMM	0.54	0.53	0.48	0.52
Neutral	0.64	0.57	0.57	0.59

3.2. Subjective evaluation

We have carried out a small listening test to assess the extent to which the transformed speech is perceived as having the intended expressivity. We also wanted to initially assess the effect of prosody modification. As the quality of prosody modification might differ between a natural and a synthesized source, we used two sentences, one taken from the voice data held out from the training sets, and one synthesized. Both were semantically unemotional. We transformed each sentence from the neutral voice towards each of the three emotional voices using each of the three methods described above, each time with and without prosody modification. In addition, we also used the corresponding recording from the expressive voices and the sentence synthesized with the target voice database, in order to verify that the targets are indeed perceived as intended. Stimuli were presented in randomized order, to five listeners, who indicated for each stimulus whether it sounded cheerful, aggressive or depressed.

The untransformed original stimuli from the expressive voices were perceived as intended, in 97% of the cases. This confirms that the intended style can be clearly perceived from the training material. The transformed samples were perceived as intended in 65% of the cases (see Table 2), which is substantially higher than chance rate (33%). Aggressive and depressed samples were perceived better than cheerful samples; frame and GMM based transformation were perceived slightly better than codebook-based transformations.

There was no global difference in recognition between versions with and without prosody modification: they were rated correct in 66% and 64% of the cases, respectively. This is surprising, as prosody usually is an important factor for emotion perception. Instead, there seems to be a clear tendency for prosody-modified transformation samples to be rated as “depressed”: 47% of the prosody-modified samples, but only 30% of the spectrally transformed stimuli without prosody modification, were rated as “depressed”. This may be due to the artifacts introduced by the FD-PSOLA based prosody modification method. There were no differences between the natural and the synthesized sentence.

Table 2: *Results of the listening test: Percent correct.*

Method	Aggressive	Cheerful	Depressed	All
Codebook	65%	55%	70%	63%
Frame	80%	45%	60%	72%
GMM	75%	55%	80%	70%
Total	73%	52%	70%	65%

4. Conclusions

In this study, we compared three methods for transforming voice quality for emotional speech synthesis. The methods were based on state-of-.the-art voice conversion algorithms for speaker identity conversion and were extended with an outlier elimination procedure and with additional smoothing steps as necessary. In objective evaluations, the three methods resulted in similar performances whereas weighted frame mapping and GMM performed slightly better in subjective evaluations.

Our future plans include the development of a hybrid algorithm that uses frame mapping and GMM methods in parallel and performs residual transformation as well. We are also planning to integrate a more robust prosody modification framework into MARY TTS based on sinusoidal modelling or harmonics/noise decomposition techniques. An extended formal subjective evaluation of the voice quality transformation methods will follow, especially on the assessment of quality and intelligibility after voice quality, prosody, and residual transformations.

5. Acknowledgements

The preparation of this paper has received funding from the DFG project PAVOQUE and from the European Community’s FP7 Programme (FP7/2007-2013) under grant agreement no. 211486 (SEMAINE).

6. References

[1] Abe, M., Nakamura, S., Shikano, K. and Kuwabara, H., “Voice conversion through vector quantization”, Proc. of the IEEE ICASSP, pp. 565-568, 1988.

[2] Stylianou, Y., Cappe, O. and Moulines, E., “Continuous probabilistic transform for voice conversion”, in IEEE Trans. on Speech and Audio Proc., vol. 6, no. 2, pp. 131-142, 1998.

[3] Kain, A. and Macon, M., "Spectral voice conversion for text-to-speech synthesis", in Proc. of the IEEE ICASSP, vol. 1, pp. 285-288, 1998.

[4] Arslan, L. M., 1999, “Speaker transformation algorithm using segmental codebooks”, in Speech Comm., vol. 28, pp. 211-226.

[5] Türk, O. and Arslan, L. M., 2006, “Robust processing techniques for voice conversion”, in Computer Speech and Language, vol. 20, pp. 441-467.

[6] Türk, O. Cross-lingual voice conversion. PhD Thesis, Bogazici University, Istanbul, Turkey, 2007.

[7] Meshabi, L., Barreaud, V. and Boeffard, O., “Comparing GMM-based speech transformation systems”, in Proc. of Interspeech, pp. 1989-1992, Antwerp, Belgium, 2007.

[8] Kang, Y., Shuang, Z., Jianhua, T., Zhang, W., and Xu, B., “A hybrid GMM and codebook mapping method for spectral conversion”, in Proc. of ACII, pp. 303-310, 2005.

[9] Chen, Y., Chu, M., Chang, E., Liu, J. and Liu, R., “Voice conversion with smoothed GMM and MAP adaptation", in Proc. of Eurospeech, pp. 2413-2416, Geneva, Switzerland, 2003.

[10] Moulines, E. and Charpentier, F., “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones”, in Speech Comm., vol. 9, pp. 453-467, 1990.

[11] Schröder, M. and Hunecke, A. “MARY TTS participation in the Blizzard Challenge 2007”, Proc. of Blizzard Challenge, Bonn, Germany, 2007.

[12] Hunecke, A. Optimal design of a speech database for unit selection synthesis. Unpublished Diploma Thesis (Diplomarbeit), Universität des Saarlandes, Saarbrücken, Germany, 2007.

[13] Schweitzer, A., Braunschweiler, N., Klankert, T., Möbius, B. and Säuberlich, B., “Restricted unlimited domain synthesis”, Proc. of Eurospeech, Geneva, Switzerland, 2003.

[14] Pammi, S., Charfuelan, M., Schröder, M. and Türk, O., “Voice Building Tool for MARY TTS”, submitted to Interspeech, Brisbane, Australia, 2008.

[15] Kawanami, H., Iwami, Y., Toda, T., Saruwatari, H. and Shikano, K., “GMM-based voice conversion applied to emotional speech synthesis”, in Proc. of Eurospeech, pp. 2401-2404, Geneva, Switzerland, 2003.