Evaluation of Expressive Speech Synthesis With Voice Conversion and Copy Resynthesis Techniques

Oytun Türk and Marc Schröder

***Abstract*—Generating expressive synthetic voices requires carefully designed databases that contain sufficient amount of expressive speech material. This paper investigates voice conversion and modification techniques to reduce database collection and processing efforts while maintaining acceptable quality and naturalness. In a factorial design, we study the relative contributions of voice quality and prosody as well as the amount of distortions introduced by the respective signal manipulation steps. The unit selection engine in our open source and modular text-to-speech (TTS) framework MARY is extended with voice quality transformation using either GMM-based prediction or vocal tract copy resynthesis. These algorithms are then cross-combined with various prosody copy resynthesis methods. The overall expressive speech generation process functions as a postprocessing step on TTS outputs to transform neutral synthetic speech into aggressive, cheerful, or depressed speech. Cross-combinations of voice quality and prosody transformation algorithms are compared in listening tests for perceived expressive style and quality. The results show that there is a tradeoff between identification and naturalness. Combined modeling of both voice quality and prosody leads to the best identification scores at the expense of lowest naturalness ratings. The fine detail of both voice quality and prosody, as preserved by the copy synthesis, did contribute to a better identification as compared to the approximate models.**

***Index Terms*—Expressive speech synthesis, prosody, voice conversion, voice quality transformation.**

I. Introduction

THE synthesis of expressive speech is a target application with potential relevance in several areas, including the dynamic generation of multimodal media content and naturalistic human–machine interaction. High-quality expressive synthetic speech can currently be obtained in niche areas only: by recording synthesis databases that contain the type of speech material needed for a specific application domain, it is possible, for example, to create a high-quality voice for simulating the voice of a poker player [1] or a shouting military officer [2]. Naturalness in this approach is obtained by applying the unit selection synthesis technology [3] to a suitable database—a database which, first, is produced with the intended expressive speaking style [4], [5] and which, second, provides good *coverage* for the target domain, thus avoiding the artifacts observed when the unit selection approach lacks suitable units to select from, such as audible breaks, or discontinuous or inappropriate prosody. By avoiding signal modification, the approach maintains the original quality of the recordings. For widespread use, however, the approach lacks flexibility: for every target domain and every intended expressive style, a new speech synthesis database would have to be recorded.

More flexibility can be obtained using a different synthesis technology, statistical-parametric synthesis using hidden Markov models (HMMs) [6]. Context-dependent Gaussian models are trained on some speech data, and then used to generate parameters to be converted into a speech waveform by means of a vocoder. For expressive synthesis, the models can either be trained on expressive speech material [7], or larger non-expressive models can be adapted to expressive speech material [8], [9]. It is also possible to generate intermediate degrees of expressions by interpolation between different styles [10], [11].

The best-sounding speech synthesis technology to date, however, remains large-corpus unit selection speech synthesis [12]. Using a very large speech corpus of, e.g., 10 hours of speech material increases the likelihood that for any given target sentence, suitable units can be found in the speech corpus, thus providing good quality synthesis. For reasons of economy and consistency, though, such very large corpora are normally produced in a default, non-expressive speaking style.

One possibility for changing the sound of a unit selection voice is to apply voice transformation technology to the generated synthetic speech. Voice transformation refers to the conversion of a given speech recording to match the acoustic characteristics of some target speech material. These desired characteristics may include the perceived speaker identity, prosody, voice quality, or accent. Vector quantization [13], weighted codebook mapping [14], and Gaussian mixture models [15], [16] have been the baseline approaches to achieve acoustic mapping and transformation in voice conversion. In addition, weighted frame mapping [17], artificial neural networks [18], and hidden Markov models [19] have also been used. Voice conversion techniques have been successfully applied for cross-language movie dubbing [20], rap singing transformation [21], expressive speech-to-speech transformation [22], and body-transmitted speech conversion for speaking aids and for noise robust speech communication [23].

In unit selection text-to-speech (TTS) domain, conventional voice transformation methods have been used for speaker identity transformation [24], expressive speech generation [25], and normalization of voice quality variations across TTS recording

Manuscript received May 08, 2009; revised December 27, 2009. Current version published June 16, 2010. This work was supported by the European Community's Seventh Framework Program (FP7/2007–2013) under Grants 211486 (SEMAINE) and 231287 (SSPNet). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yannis Stylianou.

O. Türk is with the is with Sensory, Inc., Portland, OR 97209 USA (e-mail: oytunturk@gmail.com).

M. Schröder is with the Speech Group, DFKI GmbH Language Technology Lab, D-66123 Saarbruecken, Germany (e-mail: marc.schroeder@dfki.de; schroed@dfki.de).

Digital Object Identifier 10.1109/TASL.2010.2041113

sessions [26]. Multilingual TTS is one of the most significant applications of voice transformation in TTS domain since it enables text-to-speech conversion in a speaker's voice for a language that the speaker cannot speak [27].

In order to improve naturalness in human–machine interaction, expressive speech synthesis has started to attract increasing attention in recent years. As an example, rule-based voice quality modifications were combined with decision tree-based expressive prosody prediction [28] and integrated with the well-known TTS framework MBROLA [29]. The HMM-based speech synthesis technology was used to predict and realize expressive f0 contours [30]. Rule-based emotional TTS generation [31], using phase vocoder principles for content-based emotional speech transformation [32], prosody model adaptation for expressive speech and unit selection using adapted prosody [33] have also been investigated. Copy resynthesis techniques provide a useful framework to investigate effects of individual acoustic features as well as their interactions on emotion generation and identification. Copy synthesis of prosody and vocal tract spectrum was studied to show the combined effect of spectrum and prosody at the phone level for achieving emotional synthesis [34]. In a follow-up work, prosodic parameter differences predicted at POS level for expressive speech were investigated and Gaussian transformation of prosody features was employed for neutral synthetic speech to emotional synthetic speech conversion [35].

The present paper investigates the use of voice conversion technology for expressive speech synthesis in a factorial design. Using what we expect to be clearly identifiable expressive target states (aggressive, depressed, and cheerful speech), we combine model-based and copied vocal tract and prosody and apply them to neutral speech synthesis output. This work sheds light on the following research questions.

- Which are the relative contributions of voice quality and of prosody for the identification of the expressive target states?
- To what extent is automatic voice transformation capable of modeling the expressive target state? Is the direct copy of spectral parameters from an expressive target a meaningful baseline?
- How important is the intonation *contour* as opposed to the global settings F0 level and range?
- How strongly is the perceived quality affected by the signal processing steps involved in voice quality and prosody conversion, individually and in combination?

In a previous study, we compared voice quality transformation using different voice conversion methods, and found that GMMs and weighted frame mapping showed the best performance [25]. A number of extensions to the standard GMM and codebook mapping algorithms was proposed including automatic outlier elimination based on [36], and direct frame mapping instead of codebook states with temporal smoothing [17]. The modifications were performed in a postprocessing stage after synthesizing neutral speech and then modifying it to match the target expressive style using frequency-domain pitch synchronous overlap-add (FD-PSOLA) [37] and filtering. For prosody prediction, a simple, fully statistics driven technique based on regression trees was used. Comparisons were performed in a listening test which involved forced selection of expressive style categories (neutral, aggressive, cheerful, and depressed) given a set of voice quality and prosody transformation outputs. Although this approach provided preliminary information on the performance of voice conversion algorithms in transforming neutral synthetic speech into expressive speech, a number of interesting research questions remained unattended. Specifically, as an unexpected result, the effect of prosody transformation was not significant mostly due to the artifacts in processed TTS outputs. These artifacts originated from problems in appropriate expressive prosody prediction or from large amounts of prosody modifications required or a combination of both.

In the present study, we investigate copy resynthesis of prosody as a shortcut for achieving "ideal" prosody prediction. The idea is to evaluate the performance of voice quality transformation using voice conversion techniques under the "ideally predicted" prosody conditions. The performance of GMM-based voice quality transformation is also compared to a similar copy resynthesis approach for the vocal tract. These copy resynthesis approaches are aimed at providing important clues on the amount of distortions caused by signal processing even when the target features could be predicted very accurately. In summary, this paper intends to extend the results of [25] in the following directions.

- Copy resynthesis material from recordings in target expressive styles are used for extracting "ideally predicted" target features for voice quality and prosody.
- Voice quality transformation and vocal tract copy resynthesis are cross-combined with target prosody and are evaluated in the context of expressive speech synthesis.
- Prosody modification is carried out by carefully avoiding large modification factors as well as abrupt changes over time in order to improve naturalness.
- A listening test is conducted in which not only the perceived expressive style but the quality was evaluated.
- Intended expressive styles in original neutral and expressive databases are evaluated as part of the listening test.

Section II outlines the expressive speech synthesis framework with an overview of unit selection-based speech synthesis in MARY TTS, expressive speech databases, and voice quality and prosody manipulations. Section III describes the stimuli and procedures employed in subjective evaluation of perceived expressive style and quality. Section IV provides detailed analysis and discussions of the listening test results. This paper is concluded with an overview of results and a summary of future research plans in Section V.

II. Method

Fig. 1 shows the flowchart of the expressive speech synthesis algorithm. MARY TTS Voice Building Tools are employed for creating a neutral unit selection voice [41]. Using a set of parallel utterances in the neutral style and in the target expressive style (aggressive, cheerful, or depressed), a joint-GMM is trained to transform voice quality from neutral to expressive as shown in Fig. 2. Note that the source training set consists of original recordings of the neutral voice. The synthesis stage

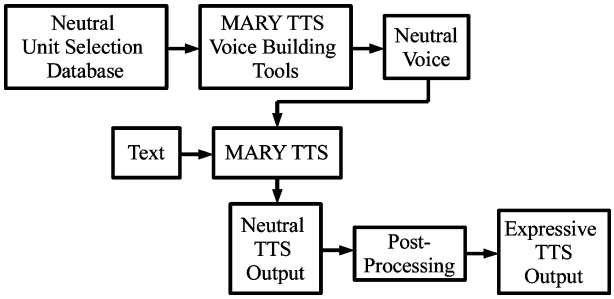


Fig. 1. Flowchart of the expressive speech synthesis algorithm.

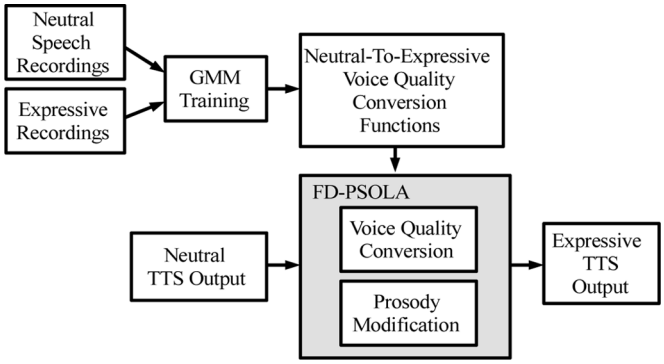


Fig. 2. Flowchart of training and postprocessing stages to transform neutral synthetic speech into expressive speech.

consists of neutral speech synthesis employing standard unit selection techniques followed by postprocessing of voice quality and prosody to obtain expressive synthetic speech output. Voice quality transformation is achieved by joint-GMM-based conversion of the spectral envelope. Pitch and duration modifications are performed using a number of approaches and cross-combined with voice quality transformation to obtain the listening test material. Implementation details of training, transformation, and modification stages are discussed in Sections II-A.

We use a German neutral speech synthesis voice, created from a database of 3000 phonetically balanced sentences ($\sim$ 5.5 hours of speech) selected from the Wikipedia [40]. A male professional actor produced the speech database in a non-expressive speaking style. We used our standard voice building tools [41] to create a unit selection voice from the data. A subset of 375 sentences was selected from the neutral recording script and produced by the same actor in a cheerful, aggressive, and depressed speaking styles. Furthermore, 25 sentences were produced in all three expressive styles but are not included in the neutral recording script. We selected four out of these 25 sentences as target material, because they require actual speech synthesis with the neutral voice but allow us to copy prosody and voice quality parameters from the expressive recordings. The sentences are random Wikipedia sentences with no obvious textual bias towards any of the expressive styles.

A. Voice Quality Conversion Training and Transformation

Conventional voice conversion algorithms for spectral envelope transformation provide a practical framework for transforming voice quality without the requirement of extraction and modification of explicit voice quality parameters. This approach works well since sufficiently detailed spectral envelope captures relevant spectral features related to voice quality including the spectral tilt, formant bandwidth, low-to-high frequency energy ratio, etc. A well-known method for spectral envelope transformation is based on GMM modeling the joint source-target line spectral frequency (LSF) space [42].

Given the three voice quality transformation methods in [25], we focus on the GMM framework for a number of reasons. First of all, GMMs result in satisfactory performance in transforming the overall voice quality in expressive speech [25], [38], [39]. GMM-based voice transformation also reduces the computational burden of text-to-speech synthesis by performing vocal tract transformation using matrix operations instead of codebook search techniques that codebook mapping and frame mapping methods employ. Additionally, the transformation function can be represented in a relatively compact manner by using a model instead of individual representatives of source and target acoustic features. In our implementation, we fit a GMM with 40 mixture components to model the acoustic mapping between the source and the target Bark-scaled LSFs using the expectation–maximization (EM) algorithm. The minimum number of EM iterations was set to 200 and the training algorithm automatically decided when to quit iterations by considering the average change in mixture means in consecutive iterations.

The training algorithm consists of feature extraction, alignment, outlier elimination, and GMM training steps. Neutral speech recordings are used as source and the corresponding expressive speech recordings as target resulting in a parallel voice conversion training approach. The recordings are labeled using an HMM-based phonetic aligner and manual correction. Linear prediction (LP) analysis is performed with a window size of 20 ms and a skip rate 10 ms using a digital pre-emphasis filter of the form $1 - 0.97z^{-1}$ and a Hanning window. The prediction order is set to 20 for a sampling rate of 16 kHz. LP coefficients are converted to LSFs which are then converted to Bark-scale. For each source LSF vector, the corresponding target LSF vector is found using the phoneme alignment information. The outliers are eliminated using the following algorithm based on [36].

- For all source and target LSF vector pairs, compute LSF distance values d_i

$$d_i = D(s_i, t_i) \tag{1}$$

  where $D(\cdot)$ is the inverse harmonic weighted LSF distance proposed in [43], s_i and t_i are the paired source (neutral) and target (expressive) LSF vectors.
- Compute μ_{d_i} and σ_{d_i}, the mean and the standard deviation of LSF distance values.
- Compute LSF distance threshold β

$$\beta = \mu_{d_i} + \alpha\sigma_{d_i} \tag{2}$$

  where $\alpha = 1.5$ works well in practice and eliminates roughly 5%–15% of the available set of LSF pairs.
- Eliminate source-target LSF pairs for which $d_i > \beta$.

In our informal evaluations, outlier elimination helped to reduce artifacts in spectral conversion since outliers in the joint source–target acoustic space are more likely to generate abrupt

changing transformation filters from one speech frame to another. The set of joint source–target LSF vector pairs which survived the elimination process are used for training a GMM to model the joint source–target LSF space.

In the transformation stage, neutral TTS output waveforms are analyzed with identical LSF extraction steps as in training. For each speech frame, the transfer function for a spectral conversion filter is obtained as follows.

- Compute target LSF estimate l^t as in [42]

$$l^t = \sum_{n=1}^{N} h_n(l)^i \left[\mu_n^t + \sum_{n}^{ts} \left(\sum_{n}^{ss} \right)^{-1} \left(l^i - \mu_n^s \right) \right] \quad (3)$$

  where l^i is the source input LSF vector, μ_n^s and μ_n^t are the mean vectors of the nth components of source and target GMMs, respectively, $\sum_n^{ss}$ is the covariance matrix of the nth component of source GMM, and $\sum_n^{ts}$ is the cross-covariance matrix of the nth component of the target-source GMM. Note that the probability of the source vector given the source GMM is given by

$$h_n(l)^i = \frac{\alpha_n N\left(l^i; \mu_n^s; \sum_{n}^{ss} \right)}{\sum_{k=1}^{N} \alpha_k N\left(l^i; \mu_k^s; \sum_{k}^{ss} \right)} \quad (4)$$

  where α_n is the weight of the nth Gaussian component as estimated by EM-based training.
- Convert l^i and l^t to LP coefficients and compute $S(w)$ and $T(w)$, the DFTs of source and target LP spectral envelopes respectively. Finally, estimate the DFT of the impulse response of the conversion filter for the source input speech frame using

$$H(w) = \frac{T(w)}{S(w)}. \quad (5)$$

In order to smooth the transformation filter given by (5), a temporal smoothing algorithm is employed in a second pass after the transformation filters for all frames are computed. Each frequency bin of $H(w)$ is temporally smoothed by using a Hanning window and corresponding frequency bins of the transformation filter from the neighboring speech frames. This smoothing step reduces discontinuities due to abrupt changes or discontinuities across frequency bins in neighboring speech frames. Choice of the number of neighboring frames plays an important role in shaping the tradeoff between voice conversion quality and similarity. Although increasing number of neighboring frames can increase output quality, it may have an adverse effect on similarity to target. After informal evaluations, the best choice for number of neighboring frames is found to be 4 in voice quality transformation tests reported in this paper. Therefore, the DFTs of the vocal tract transformation filter's impulse responses for 4 previous, 4 next, and current speech frame are used in the smoothing process using the following formula:

$$H_{smoothed}^k(w) = \sum_{i=-N}^{N} W(i) H^{k+i}(w) \quad (6)$$

where k is the speech frame index, $H^k(w)$ is the wth frequency bin of the vocal tract transformation filter, $W(\cdot)$ is a Hanning window of length $2N+1$ centered around the current frame, $N = 4$, and $H_{smoothed}^k(w)$ is the smoothed frequency bin of the transformation filter. Spectral envelope transformation is performed by multiplication of $H(w)$ with the source spectrum. Then, inverse DFT and overlap-add techniques are employed to obtain the transformed waveform in time domain.

In order to serve as a reference for voice quality transformation, we generated another set of stimuli by vocal tract copy resynthesis from the corresponding target recordings. For this purpose, phonetic labels are used in aligning the source and target speech signals. Then, linear mapping is performed within a given phoneme to determine the corresponding target LPCs for a given source speech frame. The synthesis filter is driven with these target LPCs and the source LP residual to obtain the vocal tract copy resynthesis versions.

B. Prosody Modification

We used the FD-PSOLA method for modifying pitch and timing to match target characteristics [37]. FD-PSOLA extracts linear prediction-based spectral envelope estimate using pitch-synchronous frames. Then, the short-term magnitude spectrum estimate is divided by the spectral envelope to obtain the magnitude spectrum of the LP residual. Pitch modifications are applied by compressing or expanding the spectral envelope and multiplying it with part of the LP residual (increasing f0) or copying a mirror image of the LP residual (decreasing f0). The resulting magnitude spectrum is transformed into a time-domain signal using the original phase values and inverse DFT. Duration modification is achieved by time-domain repetitions (time-scale expansion) or removals (time-scale compression) of pitch synchronous synthesis frames. In our evaluations, FD-PSOLA was driven in three ways:

- target f0 values from the corresponding target f0 contour (f0 copy re-synthesis);
- f0 scaling and shifting to match sentence mean and standard deviation of the target sentence;
- duration scaling to match target sentence duration excluding silent regions.

We have also tested copying phoneme durations directly from the target sentence recordings. However, this resulted in general loss of quality when applied to speech synthesis outputs, especially when the duration modification factor changed drastically from one phoneme to another. Therefore, we did not use any material with copied phoneme durations in the listening tests.

III. Listening Tests

In order to evaluate the quality of the expressive speech produced using the above-mentioned modification approaches, we carried out a listening test evaluation. We consider quality along two axes. First, the *expressiveness* of stimuli was verified by a forced-choice emotion identification task, in which listeners had to choose among the four categories "neutral," "gut gelaunt" (cheerful), "niedergeschlagen" (depressed), and "aggressiv" (aggressive). This tests the extent to which the different transformed versions of a sentence contain perceptual cues for the intended style. In addition, it verifies any bias

introduced by the text: the unmodified neutral synthesis version is expected to be perceived as emotionally neutral if the text does not bias listeners. Finally, this test allows us to verify the suitability of the target: the recognition rates for the original target sentences should be high if these are suitable targets for the modification task.

Second, we test the *naturalness* of stimuli, i.e., the degree to which the stimuli sound like natural human speech. This test aims to quantify the perceptual effect of signal processing artifacts introduced by the signal modification involved in the various transformation steps. We used a five-point rating scale to test naturalness, from 1 "totally unnatural" to 5 "totally natural." Listeners were instructed to ignore any emotional expressiveness in the perceived naturalness test. Including the unmodified neutral synthesis version of a sentence provides an expected upper bound for the ratings. Furthermore, the unmodified natural expressive versions of a sentence are included to represent the upper limit of the scale.

A. Stimulus Material

Four test sentences with no obvious textual bias towards one of the expressive styles were used. (1) "Ihr Sound wurde als Power-Pop beschrieben." ("Their sound was described as Power-Pop."); (2) "Prüfungen gibt es ab dem sechsten Kyu." ("Examination starts with the sixth Kyu."); (3) "Sie riechen intensiv nach Pfeffer." ("They smell intensely like pepper."); (4) "Bei Wagram wurde er verwundet." ("He was wounded close to Wagram."). For each of the sentences, the following versions were used as stimuli:

- one neutral synthesis version, unmodified;
- for each expressive target style,
  - the original spoken expressive target recording;
  - 14 transformed versions of the neutral synthesis output, combining three vocal tract settings (no modification, copy, GMM-based prediction), with five prosody settings (no modification, F0 contour copy, F0 mean and standard deviation adaptation, F0 contour copy and sentence duration adaptation, F0 mean and standard deviation adaptation combined with sentence duration adaptation).[1]

Therefore, a total of 46 versions of each sentence were included in the test −42 transformed versions and 4 original versions. For all four sentences, this yields 168 transformed and 16 original stimuli.

B. Test Procedure

In order to avoid exceedingly long listening tests, the stimuli were split into two sets. Each set included the 16 original stimuli as well as half of the transformed stimuli, selected pseudo-randomly such that each transformation condition was equally represented in each set. The full set of original stimuli was retained in both sets so that all listeners encounter all originals of each sentence, in order to serve as reference points for calibrating the perceptual scale on which the ratings are made. This resulted in two sets A and B, of 100 stimuli each.

[1]We ommitted the version combining no vocal tract modification with no prosody modification, which is identical to the unmodified neutral synthesis stimulus.

TABLE I
EMOTION IDENTIFICATION RATES (ORIGINAL STIMULUS)

		Identified			
		Neutral	**Cheerful**	**Depressed**	**Aggressive**
Intended	**Neutral**	**94.8**	3.1	2.1	0.0
	Cheerful	20.8	**72.9**	0.0	6.3
	Depressed	1.0	0.0	**98.0**	1.0
	Aggressive	4.2	2.1	0.0	**93.7**

Emotion identification rates are shown in percent. Original stimulus consists of neutral speech synthesis outputs and original expressive recordings.

Each subject either performed the expression identification task on set A and the perceived naturalness task on set B or vice versa. The emotion identification task was always carried out first, in order to allow the listeners to get acquainted with the material before making quality judgments. For this task, the subjects selected among neutral, cheerful, depressed, and aggressive options. 24 subjects (11 male, 13 female, between ages 21–54 with mean age of 28.0 years) participated in the test. All subjects were native German speakers with no known hearing problems, eight of whom were research scientists in language and speech technology. The test was conducted using high quality headphones connected to a laptop in a quiet environment. The emotion identification task took around 11 minutes and the perceived naturalness task around 8 minutes to be completed.

IV. RESULTS AND DISCUSSION

A. Emotion Identification Task

Table I shows the confusion matrix using the original stimuli (neutral synthesis outputs and original expressive recordings). All categories are predominantly perceived as intended, with identification rates above 90% for neutral, depressed, and aggressive. This pattern is consistent with the expectation that identification should be easy for acted material with a small number of categories [44]. The lower identification rate of 72.9% for cheerful is due to a confusion with neutral, which suggests that the cheerful style as rendered by our actor on the target sentences was not quite as typical for a cheerful style as his renditions of other styles on the target sentences. These values serve as reference for interpreting the identification scores of transformed stimuli below. The reader should also note that we used the abbreviations in Table II to represent different vocal tract and prosody manipulation settings; "+" signs indicate that two or more manipulations are applied together.

Table III shows correct emotion identification rates for different combinations of vocal tract, pitch, and duration manipulations. We have performed four-way analysis of variance (ANOVA) to examine the effect of target emotion ("cheerful," "depressed," "aggressive"), vocal tract manipulation setting ("None," "Copy," "GMM"), prosody manipulation setting ("None," "f0 copy," "sentence f0 mean/standard deviation," "f0 copy + duration," "sentence f0 mean/standard deviation + duration ") and sentence type (1 of 4 sentences) on these emotion identification rates. ANOVA results show that target

TABLE II
ABBREVIATIONS FOR DIFFERENT VOCAL TRACT AND PROSODY MANIPULATIONS

Abbreviation	Processing
v1	Vocal tract copy re-synthesis from target
v2	Vocal tract transformation using GMM
p1	Pitch contour copy re-synthesis from target
p2	Pitch contour transformation to match the mean and the variance of the corresponding target recording´s pitch contour
d1	Duration transformation to match target sentence duration (excluding silence)

TABLE III
EMOTION IDENTIFICATION RATES IN DETAIL

		Percent identified correct												
		Ori-ginal	v1	v2	p1	p2	v1+p1	v1+p1+d1	v1+p2	v1+p2+d1	v2+p1	v2+p1+d1	v2+p2	v2+p2+d1
Trans-formed	Neutral-to-cheerful	72.9	12.5	2.1	70.8	36.2	81.2	79.2	45.8	47.9	77.1	85.4	47.9	47.9
	Neutral-to-depressed	98.0	29.8	27.7	68.1	43.8	91.7	93.8	74.5	74.5	72.9	89.6	64.6	79.2
	Neutral-to-aggressive	93.7	39.6	10.4	6.4	0.0	70.2	62.5	66.7	66.0	47.9	43.8	25.3	35.4
	Average	89.8	27.3	13.3	48.6	26.6	81.1	78.5	62.2	62.7	66.0	72.9	46.2	54.2

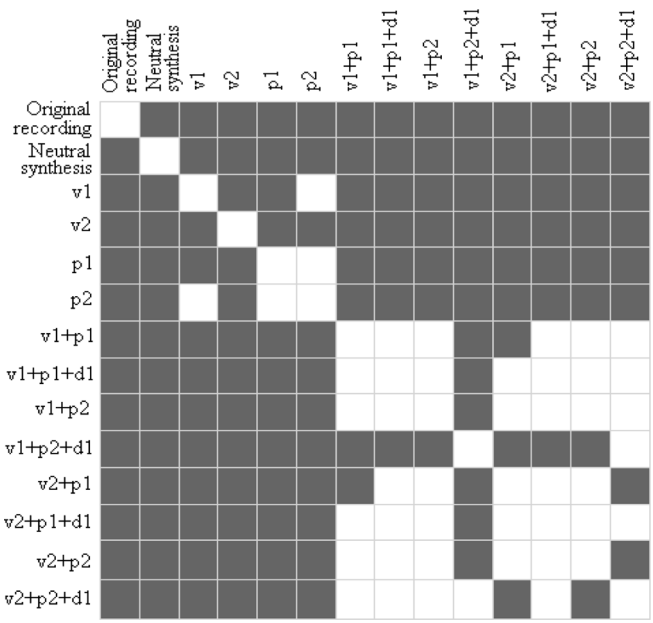
Emotion identification rates are shown in percent for different combinations of vocal tract, pitch and duration manipulations.

emotion ($p < 0.01$, F- ratio $= 109.0$), vocal tract manipulation setting ($p < 0.01$, F- ratio $= 54.89$), and prosody manipulation setting ($p < 0.01$, F-ratio $= 97.16$) had significant effects alone. Sentence type had a smaller and nonsignificant effect for a confidence level of 99.0% ($p = 0.0132$, F-ratio $= 3.59$). In addition to these main effects, interesting two-way interactions included "target emotion and vocal tract manipulation setting" ($p < 0.01$, F-ratio $= 30.59$) and "target emotion and prosody manipulation setting" ($p < 0.01$, F-ratio $= 7.17$).

Considering significant main effects found in ANOVA and the corresponding average identification scores in Table III, we observe that neither vocal tract transformation (v1 or v2) nor prosody modification (p1 or p2) alone result in equivalent performance as compared to identification rates for original recordings. The identification of emotion is most successful when copy synthesis of both vocal tract (v1) and prosody (p1 and $p1 + d1$) is combined. Approximate modeling (v2 and p2) goes part of the way.

There are clear differences between the emotions. Whereas depressed and cheerful speech are identified approximately to the best extent that could be expected from the original recordings' recognition scores (Table I), identification of aggressive speech is somewhat lower. Apparently, some perceptual cues that lead to the identification of aggressiveness in the original recordings are not modeled in the transformed stimuli. It is conceivable that this cue is related to energy, which was not modeled in this experiment. Regarding the relative importance of prosody for emotion identification, the results suggest that

TABLE IV
WILCOXON SIGNED RANK TEST RESULTS FOR PERCEIVED NATURALNESS RATINGS. PERCEIVED NATURALNESS RATINGS OBTAINED FOR EACH PAIR ARE COMPARED USING WILCOXON SIGNED RANK TEST FOR A CONFIDENCE LEVEL OF 99.0%. GRAY AREAS INDICATE STATISTICALLY SIGNIFICANT RATING DIFFERENCES FOR THE CORRESPONDING METHOD PAIR



"cheerful" and "depressed" were recognized from the prosody. When the exact shape of the pitch contour (p1) was available,

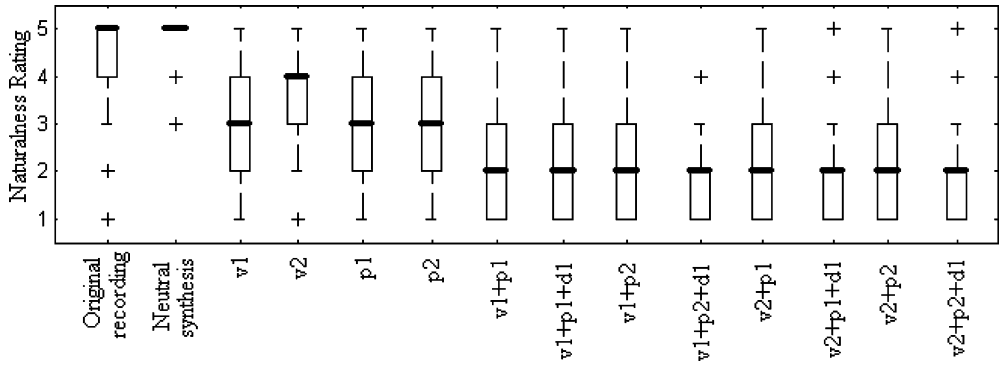


Fig. 3. Perceived naturalness ratings for original recordings, synthesis outputs, and synthesis outputs postprocessed with different methods. Thick lines mark the median for each perceived naturalness rating distribution.

TABLE V
AVERAGE PERCEIVED NATURALNESS RATINGS

Average rating													
Original recording	Neutral synthesis	v1	v2	p1	p2	v1+p1	v1+p1+d1	v1+p2	v1+p2+d1	v2+p1	v2+p1+d1	v2+p2	v2+p2+d1
4.6	4.8	3.1	3.8	2.9	2.9	2.0	2.1	2.1	1.9	2.2	2.1	2.2	2.0

recognition was better than based on average f0 level and range alone (p2). Applying duration scaling to match the average target sentence duration in addition to vocal tract and pitch manipulations ($v1{+}p1{+}d1, v1{+}p2{+}d1, v2{+}p1{+}d1, v2{+}p2{+}d1$) does not increase identification rates significantly. The "aggressive" target style was not recognized from prosody alone.

The perceptual contribution of voice quality appears to be roughly complementary to prosody. By itself (v1 and v2), it does not lead to the recognition of "cheerful" speech, and to a perception of "depressed" speech close to chance level. For "aggressive," the copy resynthesis of voice quality (v1) is recognized above chance level. When combining copied voice quality and prosody ($v1 + p1$), a roughly additive effect for the contributions of prosody and voice quality can be observed for "cheerful" and "depressed" speech; for "aggressive" speech; however, the recognition is rather clearly higher than the sum of individual contributions, suggesting a configurational effect: aggressive speech seems to be recognized from the combination of prosodic and voice quality cues in this experiment.

The GMM-based vocal tract transformation (v2) does not lead to correct identification of the intended speaking style when used in isolation, but when combined with f0 copy resynthesis ($v2{+}p1$) and duration modification ($v2{+}p1{+}d1$), it goes part of the way towards the recognition rates achieved with a full copy of voice quality and prosody ($v1 + p1$ and $v1 + p1 + d1$, respectively).

B. Perceived Naturalness Rating Task

In analyzing the naturalness ratings, we follow the evaluation framework of the Blizzard Challenge [12], the annual state-of-the-art TTS joint evaluation task. Following what is regarded as best practice there, we use median values (Fig. 3) and the results of pairwise Wilcoxon signed rank tests [45] (Table IV) for discussing the results and their significance. In addition, and for information only, we also show averages of perceived naturalness ratings (Table V).

The perceived naturalness ratings show a clear pattern of the amount of degradations introduced by the different transformations. As expected, original expressive recordings and neutral synthesis results have very high naturalness ratings. The transformed stimuli with the highest ratings are the vocal-tract-only transformations using GMMs (v2). Their ratings are significantly higher than all the other transformed stimuli's ratings. In particular, they are rated better than the direct copy synthesis (v1); this may be due to the fact that the GMM-based prediction method performs smoothing of the transformation filter. A second group is formed by the other single-transformation stimuli, vocal-tract copy synthesis (v1), F0 contour copy synthesis (p1), and F0 mean and standard deviation transformation (p2). Clearly worse than these, again, is the group of stimuli which show both vocal tract and prosody transformations.

We can therefore say that model-based vocal tract transformation appears to introduce relatively limited artifacts when it is the only transformation applied; the artifacts introduced by prosody modification in isolation are also limited. However, when both methods are combined, we observe an additive effect of the distortions, leading to a considerably worse rating. This may at least in part be due to the fact that the steps of unit concatenation, voice quality transformation, and prosody modification are all performed separately. Furthermore, voice quality transformation is done in LSF space, whereas prosody modification is done in Fourier-transformed frequency space. If the repeated analysis-synthesis steps each add their own distortions, which seems likely, then it is not surprising that the combination of both exhibit more distortions than any of them in isolation.

V. CONCLUSION

In this paper, we have investigated the joint transformation of vocal tract and prosodic features in the context of expressive unit

selection speech synthesis. For voice quality transformation, we have used two methods: GMMs modeling the joint source-target LSF space, as well as direct copy synthesis from an expressive target. An extra smoothing step for transformation filters have been employed to improve naturalness in GMM-based transformation. For prosody transformation, we have used FD-PSOLA to either copy the actual target contour or modify only mean and standard deviation of pitch, with or without a global duration adaptation. In a listening test evaluation, we have shown that there is a tension between identification and naturalness: combined modeling of both voice quality and prosody leads to the best identification scores but to the lowest naturalness ratings.

The fine detail of both voice quality and prosody, as preserved by the copy synthesis, did contribute to a better identification as compared to the approximate models. This suggests it is worth the effort to create high-quality predictive models for the expressive targets, that go beyond global mean and standard deviation parameters for characterizing expressive prosody.

In the future, we will experiment with an integrated approach where unit concatenation, voice quality transformation and prosody modification are all performed in a single parameter space. The fact that harmonics-plus-noise modeling (HNM) [46] seems effective in reducing distortions in pitch smoothing for unit concatenation compared to TD-PSOLA [47] suggests that HNM may also lead to fewer distortions than FD-PSOLA when imposing an expressive pitch contour. Furthermore, if voice quality transformation and prosody modification can both be carried out on an HNM parameter representation, it may be that the distortions introduced by combining both kinds of transformation are less severe than in the current, separate modification approach.

Sample outputs are available at: http://www.dfki.de/~schroed/samples/TurkSchroeder2010/

ACKNOWLEDGMENT

The authors would like to thank all listeners and to the student assistants A. Klepp and J. Schmidt for their contributions in the listening tests.

REFERENCES

[1] P. Gebhard, M. Schröder, M. Charfuelan, C. Endres, M. Kipp, S. Pammi, M. Rumpler, and O. Türk, "IDEAS4Games: Building expressive virtual characters for computer games," in *Proc. IVA 2008*, Tokyo, Japan, pp. 426–440.

[2] W. L. Johnson, S. S. Narayanan, R. Whitney, R. Das, M. L. Bulut, and C. LaBore, "Limited domain synthesis of expressive military speech for animated characters," in *Proc. ICSLP 2002*, Denver, CO.

[3] A. W. Black and N. Campbell, "Optimising selection of units from speech databases for concatenative synthesis," in *Proc. Eurospeech*, Madrid, Spain, 1995, pp. 581–584.

[4] A. W. Black, "Unit selection and emotional speech," in *Proc. Eurospeech*, Geneva, Switzerland, 2003.

[5] A. Iida and N. Campbell, "Speech database design for aconcatenative text-to-speech synthesis system for individuals with communication disorders," *Int. J. Speech Technol.*, vol. 6, pp. 379–392, 2003.

[6] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, Budapest, Hungary, 1999.

[7] R. Tsuzuki, H. Zen, K. Tokuda, T. Kitamura, M. Bulut, and S. S. Narayanan, "Constructing emotional speech synthesizers with limited speech database," in *Proc. ICSLP*, Jeju, Korea, 2004.

[8] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano, "Model adaptation approach to speech synthesis with diverse voices and styles," in *Proc. ICASSP*, Honolulu, Hawaii, 2007, pp. 1233–1236.

[9] M. Tachibana, S. Izawa, T. Nose, and T. Kobayashi, "Speaker and style adaptation using average voice model for style control in HMM-based speech synthesis," in *Proc. ICASSP*, Las Vegas, NV, pp. 4633–4636.

[10] K. Miyanaga, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based speech synthesis," in *Proc. ICSLP*, Jeju, Korea, 2004.

[11] T. Nose, J. Yamagishi, and T. Kobayashi, "A style control technique for speech synthesis using multiple regression HSMM," in *Proc. INTERSPEECH 2006*, Pittsburgh, PA, USA.

[12] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Proc. Blizzard Challenge 2008*, Brisbane, Australia.

[13] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. IEEE ICASSP*, 1988, pp. 565–568.

[14] L. M. Arslan, "Speaker transformation algorithm using segmental codebooks," *Speech Commun.*, vol. 28, pp. 211–226, 1999.

[15] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[16] A. B. Kain, "High resolution voice transformation," Ph.D. dissertation, OGI School of Sci. and Eng., Oregon Health and Sci. Univ., Beaverton, 2001.

[17] O. Türk, "Cross-lingual voice conversion," Ph.D. dissertation, Boğaziçi Univ., Istanbul, Turkey, 2007.

[18] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proc. IEEE ICASSP*, Taipei, Taiwan, Apr. 2009.

[19] E.-K. Kim, S. Lee, and Y.-H. Oh, "Hidden markov model based voice conversion using dynamic characteristics of speaker," in *Proc. Eurospeech*, 1997, pp. 2519–2522.

[20] O. Türk and L. M. Arslan, "Subband based voice conversion," in *Proc. ICSLP*, Denver, CO, Sep. 2002, vol. 1, pp. 289–292.

[21] O. Türk, O. Büyük, A. Haznedaroglu, and L. M. Arslan, "Application of voice conversion for cross-language rap singing transformation," in *Proc. IEEE ICASSP*, Taipei, Taiwan, Apr. 2009.

[22] Z. Inanoglu and S. J. Young, "A system for transforming the emotion in speech: Combining data-driven conversion techniques for prosody and voice quality," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 27–31, 2007.

[23] T. Toda, K. Nakamura, H. Sekimoto, and K. Shikano, "Voice conversion for various types of body transmitted speech," in *Proc. IEEE ICASSP*, Taipei, Taiwan, Apr. 2009.

[24] T. En-Najjary, O. Rosec, and T. Chonavel, "A voice conversion method based on joint pitch and spectral envelope transformation," in *Proc. 8th Int. Conf. Spoken Lang. Process.*, Jeju, Korea, 2004.

[25] O. Türk and M. Schröder, "A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 2282–2285.

[26] Y. Stylianou, "Assessment and correction of voice quality variabilities in large speech databases for concatenative speech synthesis," in *Proc. IEEE ICASSP*, Phoenix, AZ, 1999.

[27] J. Latorre, K. Iwano, and S. Furui, "Polyglot synthesis using amixture of monolingual corpora," in *Proc. IEEE ICASSP*, 2005, vol. 1, pp. 1–4.

[28] F. Tesser, P. Cosi, C. Drioli, and G. Tisato, "Emotional festival-Mbrola TTS synthesis," in *Proc. Interspeech*, Lisbon, Portugal, 2005.

[29] T. Dutoit and H. Leich, "Text-to-speech synthesis based on aMBE re-synthesis of segments database," *Speech Commun.*, vol. 13, pp. 435–440.

[30] K. Hirose, "Improvement in corpus-based generation of f0 contours using generation process model for emotional speech synthesis," in *Proc. Interspeech*, 2004, pp. 1349–1352.

[31] E. Zovato, A. Pacchiotti, S. Quazza, and S. Sandri, "Towards emotional speech synthesis: A rule based approach," in *Proc. 5th ISCA Speech Synth. Workshop*, Jun. 2004.

[32] G. Beller and X. Rodet, "Content-based transformation of the expressivity in speech," in *Proc. 16th Int. Congr. Phonetic Sci.*, Saarbrücken, Germany, Aug. 2007, pp. 2157–2160.

[33] D. Jiang, W. Zhang, L. Shen, and L. Cai, "Prosody analysis and modeling for emotional speech synthesis," in *Proc. IEEE ICASSP*, Mar. 2005, vol. 1, pp. 281–284.

[34] M. Bulut, C. Busso, S. Yildirim, A. Kazemzadeh, M. C. Lee, S. Lee, and S. Narayanan, "Investigating the role of phoneme-level modifications in emotional speech resynthesis," in *Proc. Interspeech*, Lisbon, Portugal, 2005.

[35] M. Bulut, S. Lee, and S. Narayanan, "A statistical approach for modeling prosody features using POS tags for emotional speech synthesis," in *Proc. IEEE ICASSP*, Honolulu, HI, Apr. 2007, vol. 4, pp. 1237–1240.
[36] O. Türk and L. M. Arslan, "Robust processing techniques for voice conversion," *Comput. Speech Lang.*, vol. 20, pp. 441–467, 2006.
[37] E. Moulines and W. Verhelst, "Time-domain and frequency-domain techniques for prosodic modification of speech," in *Speech Coding and Synthesis*, Kleijn and Paliwal, Eds. Amsterdam, The Netherlands: Elsevier, 1995, pp. 519–555.
[38] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis," in *Proc. Eurospeech*, 2003, pp. 2401–2404.
[39] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken English," *Speech Commun.*, vol. 51, no. 3, pp. 268–283, Mar. 2009.
[40] A. Hunecke, "Optimal design of a speech database for unit selection synthesis," Diploma thesis, Univ. des Saarlandes, Saarbrücken, Germany.
[41] M. Schröder, M. Charfuelan, S. Pammi, and O. Türk, "The MARY TTS entry in the Blizzard Challenge 2008," in *Proc. Blizzard Challenge 2008*, Brisbane, Australia.
[42] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. IEEE ICASSP*, 1998, vol. 1, pp. 285–288.
[43] R. Laroia, N. Phamdo, and N. Farvardin, "Robust and efficient quantization of speech LSP parameters using structured vector quantizers," in *Proc. IEEE ICASSP*, 1991, pp. 641–644.
[44] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, Lisbon, Portugal, 2005.
[45] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull. 1*, pp. 80–83, 1945.
[46] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods for speech and speaker modification," Ph.D. dissertation, Dept. Signal, Ecole Nationale Superieure des Telecomm. , ENST-Telecom Paris, Paris, France, 1996.
[47] A. Syrdal, Y. Stylianou, L. Garrison, A. Conkie, and J. Schroeter, "TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis," in *Proc. IEEE ICASSP*, Seattle, WA, 1998, pp. 273–276.

Oytun Türk received the B.S. degree, M.S., and Ph.D. degrees in electrical and electronics engineering from Bogazici University, Istanbul, Turkey, in 2000, 2003, and 2007, respectively.

In 2000, he cofounded Sestek A.S., where he worked as a Lead Researcher in commercial voice conversion technology. He has managed research projects in voice conversion and speech therapy, and participated in the European Union's Sixth Framework Network of Excellence SIMILAR. Since 2007, he has been working on adaptation/interpolation of expressive style, prosody, and speaker identity in open-source MARY TTS platform (http://mary.dfki.de). He has authored/coauthored over 25 publications in speech and audio processing, holds two patents in voice conversion, and serves as a reviewer in major speech and signal processing journals.

Marc Schröder received the Maîtrise in language science from Université Grenoble 3, Grenoble, France, in 1998 and the Ph.D. degree in phonetics from Saarland University, Saarbruecken, Germany, in 2003.

He is a Senior Researcher at DFKI and the leader of the DFKI Speech Group, DFKI GmbH Language Technology Lab, Saarbruecken. Since 1998, he has been responsible for building up technology and research in TTS at DFKI. Within the FP6 NoE HUMAINE, he has built up the scientific portal (http://emotion-research.net). He is editor of the W3C Emotion Markup Language specification, Coordinator of the FP7 STReP SEMAINE, and project leader of the national-funded basic research project PAVOQUE. He is an author of more than 50 scientific publications and PC member in many conferences and workshops.

Dr. Schröder won the Grand Prize for the best IST project website in 2006.