A NOVEL METHOD FOR PROSODY PREDICTION IN VOICE CONVERSION

Elina E. Helander

Tampere University of Technology
Institute of Signal Processing
elina.helander@tut.fi

Jani Nurminen

Nokia Research Center
Multimedia Technologies laboratory
jani.k.nurminen@nokia.com

ABSTRACT

Most of the published voice conversion schemes do not consider detailed prosody modeling but only control the F0 level and range. However, the detailed prosody can also carry a significant amount of speaker identity related information. This paper introduces a new method for converting the prosody in voice conversion. A syllable-based prosodic codebook is used to predict the converted F0 using not only the source contour but also linguistic information and segmental durations. The selection of the most suitable target contour is carried out using a trained classification and regression tree. The F0 contours in the codebook are represented in a transformed domain which allows compression and fast comparison. The performance of the prosodic conversion is evaluated in a real voice conversion system. The results indicate a significant improvement in speaker identity and naturalness when compared to GMM (Gaussian mixture model) based pitch prediction approach.

Index Terms— Voice conversion, prosody conversion, prosodic codebook

1. INTRODUCTION

Voice conversion (VC) has been an active research topic during the past two decades (e.g. [1], [2], [3], [4], [5]), although commercial usage of the technology has not yet been popular. The term voice conversion refers to the modification of the speech of one speaker (source) to sound as if it was uttered by another speaker (target), while maintaining the lexical content of uttered speech. Even though short-term spectral conversion has gained a lot of interest, there has been only a little consideration of converting the prosodic features like F0 movements and speaking rhythm. In this paper, we use the term prosody conversion to refer to F0 and durational modifications. The main focus is on F0 modification.

One potential application for voice conversion techniques is usage in a text-to-speech (TTS) system. In this context, voice conversion could be used for easy generation of TTS voices. The original TTS voice recordings have usually been spoken with flat prosody in order to better enable smooth concatenation of units, and this creates challenges also for the voice conversion techniques if the TTS voice is used as a source voice.

Prosody is affected not only by the content of the sentence but also speaker-specific variations. For the same sentence, different people generally produce different prosody. In addition, the same person can also utter the same sentence with quite different prosodies if her/his speech is recorded several times. Due to these reasons, the goal in prosodic conversion can be declared to be the generation of ”believable” prosody, i.e. prosody that the target speaker *could* use in a certain context.

Many of the publications in the area of voice conversion neglect detailed prosodic modeling. The most popular method is to adjust the F0 level and range of every F0 measurement point f_s as

$$f_t = \frac{f_s - \mu_s}{\sigma_s} \cdot \sigma_t + \mu_t \qquad (1)$$

where μ_s, σ_s, μ_t, and σ_t represent mean and standard deviation (std) of the F0 values for the source and the target, respectively. In the paper, this approach is referred to as the MS method. The MS method retains the shapes of source F0 contours and cannot model local changes in F0. A more sophisticated mapping function can be obtained by representing pitch values with GMMs (Gaussian mixture models). This approach actually leads to a weighted sum of different mapping functions. The GMM based pitch prediction [6] is the method that our approach is examined against in the listening tests in Section 4.

There are not many proposals in the literature for converting F0 contours or other prosodic features in more detail. Chapell et al. [7] described three F0 prediction methods: the MS method as in (1), a similar approach but using a cubic fit to the data, and an utterance level codebook. These three methods and a voiced contour codebook were evaluated in [8]. A closely-related idea of voiced contour codebook was presented previously in [9] but instead of picking only one contour like in [8], the resulting contour was formed by using a weighted average of all contours according to the distance between a codebook source contour and the voiced contour to be transformed. In [10], sentence-initial high and sentence-final low F0 values as well as hand-marked accents were modeled using separate means and standard deviations.

In the method introduced in this paper, a syllable-level codebook containing paired source and target F0 contours is built. The F0 information is compressed using discrete cosine transform (DCT) coefficients that enable fast comparison and make it possible to avoid the use of dynamic time warping (DTW) based techniques for alignment. The codebook source contours that are close enough to the source contour under conversion (referred to as SCUC) are regarded as candidates and the final decision is made based on a CART (classification and regression tree) that is trained on linguistic and durational features.

The F0 contours stored in the codebook are actual contours estimated directly from the speech data without any clustering or averaging of contours. In addition, only one target contour is chosen as the output in order to avoid the risk of obtaining a flat contour as a result of weighted averaging. The linguistic context is taken into account in the selection to avoid producing linguistically incorrect contours.

The paper is organized as follows. In Section 2, the proposed prosody conversion approach is introduced on a general level. In

1-4244-0728-1/07/$20.00 ©2007 IEEE

Section 3, the method is described in detail. The experimental results obtained in a listening test are presented in Section 4 whereas Section 5 concludes the work.

2. GENERAL DESCRIPTION OF THE METHOD

The prosody models used in TTS systems are usually trained using a very large database (consisting of e.g. 10 hours of speech). In voice conversion, the size of the training set has to be much smaller (only a few minutes), e.g. to enable online training for the users. Consequently, the aim in this work is not to build a sophisticated prosody model but only to capture some major tendencies from the data.

We propose to model prosody at the syllable level. Syllable is a linguistically determined unit but it can be also considered prosodically justified. Prosodic events take place in synchrony with syllables or groups of syllables. For example, the tone sequence theory on intonation modeling concentrates on F0 movements on syllables. The syllable level also seems to be quite robust on labeling errors. According to the sonority principle in English [11] that also applies to many other languages, the voicing should be continuous in the syllable. This issue is beneficial from the viewpoint of obtaining meaningful contours for the codebook.

The prosodic codebook is generated by first collecting syllable-aligned F0 contours using parallel training sentences from the source and the target speakers. The F0 contours are transformed into DCT coefficients and this information is stored in the codebook as vector pairs. Linguistic and durational information is also stored for each entry. As a second part of the training process, a CART is trained using the codebook as the training data. The role of the CART is to help in the selection process.

During conversion, the SCUC is transformed into DCT domain and compared to the codebook keys (source contours). The codebook entries whose source contours are close enough to the SCUC are chosen as candidates for the final selection that is performed using the CART. The target contour of the selected entry is IDCT (inverse DCT) transformed and given a F0 level value from the mean level of the current syllable predicted with the MS method. These steps are described in more detail in the following subsections.

3. DETAILED DESCRIPTION OF THE METHOD

3.1. Codebook generation

Using parallel corpora from the source and the target speakers with boundary labels and linguistic descriptions, the syllable-length F0 contours are obtained using a pitch estimation algorithm. The resulting source and target contours of each syllable can further be smoothed, if necessary, and possible F0 outliers at the syllable boundaries can be removed. The syllables containing voiced contours that have a too short duration for meaningful contour representation are discarded. For all the other syllables, the process is continued by applying DCT on the contours. M first DCT coefficients of the source and the target are stored, denoted as $\mathbf{s}_k$ and $\mathbf{t}_k$ for the syllable k, respectively. The first coefficient does not have to be stored as it can be set to zero since it represents the bias (F0 level) that is handled separately. Due to the truncation or zero-padding to the length M, the coefficients are normalized by the factor $\sqrt{N}$, where N denotes the length of the original contour.

In addition to the DCT domain contours, simple linguistic information and duration features are stored in the codebook for each entry. This information can be obtained from almost any TTS system, without training specific models. We have decided to use features that have also been popularly used in data-driven prosody generation techniques: lexical stress, local position in the word {*initial, mid, final, monosyllabic*}, global position in the phrase {*initial, final, first in a prosodic phrase* (predicted using simple punctuation rules), *none*}, Van Santen-Hirschberg classification for onset as well as coda {*unvoiced, voiced but no sonorants, sonorant*} and the type of the word the syllable belongs to {*content, function*}. In addition to the linguistic information related to a specific syllable, the information related to the previous and the next syllable can also be taken into account. As the duration related features, the total duration of the syllable for the source and the target, respectively, and the duration of the voiced contour of the source and the target, respectively, are stored in the codebook for each entry.

3.2. CART training

In the training of the CART, the design goal is to build a tree that can output an optimality score based on the linguistic and durational similarity. The process begins with the generation of the training data. As a preliminary step, two distance matrices are computed based on the codebook. The elements of a source-side distance matrix **A** are computed as

$$a_{kj} = \sum_{m=0}^{M-1} (s(m)_k - s(m)_j) \qquad k, j = 1, 2, \ldots, K \qquad (2)$$

where K denotes the number of syllables in the codebook. As can be seen from the equation, the element a_{jk} gives the distance between the source contours j and k. A similar distance matrix **B** is computed for the target contours.

The training data is formed from the codebook data as follows. All the entries in the codebook are taken into consideration, one by one. For the entry j, this means that the source contour of this entry is compared against the source contours of the other entries based on the elements of matrix **A** from a_{j1} to a_{jK} except for a_{jj}. If a_{jk} is below a threshold, i.e. $a_{jk} < \delta_j$, the corresponding entry k is considered a potential candidate for being a good substitute for the entry j. The threshold δ_l is made adaptive on the source contour of the entry l in such a way that a p% deviation from the closest match is allowed in terms of contour distance. For each potential candidate, the corresponding target distance b_{jk} is obtained. Based on b_{jk}, the entry k is considered either a ”possibly optimal”, a ”neutral” or a ”non-optimal” candidate as an substitute for the entry j.

The codebook entries below an experimentally tuned threshold β_o are considered possibly optimal choices and the entries above the threshold β_n represent the non-optimal case. The neutral cases falling between these thresholds are not used in the training since they fall into an uncertain region. For the possibly optimal and non-optimal entries, the linguistic information is compared against the linguistic information of the entry j, resulting in a binary vector. In the binary vector, each zero means that there was a match in the corresponding feature (for example 0 if both are monosyllabic), while the value 1 means that the corresponding features were not the same. In addition to the binary distances, the absolute differences of the syllable durations and the voiced durations are also computed and stored for usage as the training data. After repeating the above procedure for all the entries in the codebook, the generated training data consists of a reasonably large amount of data from the two classes (”possibly optimal” and ”non-optimal”) with the corresponding linguistic and durational information.

The training of the CART aims at finding which features are important in the final candidate selection. There can be many codebook entries that have quite similar source contours but clearly different target contours, and thus finding out how much the duration and the context affect the situation is important. In the CART training, we have used a CART with gini impurity measure [12]. The CART was pruned according to the results of 10-fold cross-validation in order to prevent over-fitting and the terminal nodes were pruned if they ended up having only small number of observations.

3.3. Prediction of F0 contours

The conversion process starts from syllable boundary detection. The DCT coefficients representing the syllable-length F0 source contours are calculated and zero-padded or truncated to length M, taking into account the normalization as described in Section 3.1. For the syllables that do not contain sufficiently many F0 values for obtaining a meaningful contour representation, the MS method is used in the F0 prediction. Otherwise, the process starts similarly as in the training: Some codebook entries become candidates based on the small-enough difference between the source contours, computed as in (2). The threshold for accepting candidates is determined based on the minimum difference, allowing a p% deviation. If the threshold becomes too high, the MS method is used. In all other cases, the linguistic information between the SCUC and the candidates is matched and the absolute differences in the syllable duration as well as in the voiced contour duration are calculated. This information is used as an input to the CART, and the candidate leading to the tree node producing the highest probability for the possibly optimal class is chosen as the selected codebook entry. If there are two or more candidates producing the highest probability, the entry candidate whose source contour's difference to SCUC is the smallest is selected.

After selecting the most appropriate entry from the codebook as described above, the final contour is produced by taking the IDCT of the corresponding target contour. The length in the DCT domain is zero-padded or truncated to match the length of the SCUC, together with appropriate weighting in order to obtain a contour having the correct length (the possible duration change is handled separately). Next, the F0 level is added to the contour. If the original F0 contour is continuous across the boundary of two syllables k and k+1, the converted contours are also made continuous by adding a bias value to syllable k+1. The bias is determined as the difference between the last point of syllable k and the first point of syllable k+1. Since this can result in major changes in F0 std calculated together for the two syllables, the std is scaled back to the level where it was before the change. In addition, the F0 level is also set again for these syllables, now calculated together.

Conventionally, durations are modeled using a simple utterance level scaling. We propose to apply syllable-level scaling using regression coefficients calculated from all the source and target syllable durations. This results in more detailed modifications than the utterance-level scaling. Alternatively, the duration scaling ratios can be predicted by building a CART using the linguistic features. A third alternative is to use directly the target syllable duration that corresponds to the chosen index.

4. EXPERIMENTAL RESULTS

4.1. How to evaluate prosody conversion?

There is no generally accepted objective measure for evaluating prosody conversion. Since there are no strictly right or wrong F0 contours for the target speaker, the goal should be to achieve acceptable and believable prosody. In the literature, no evaluations were done in [13] nor in [7]. In [8], the converted pitch was transplanted to the real target utterance using dynamic time warping. Although the intention is to prevent the spectral conversion from affecting the result, in our experiments this kind of evaluation did not reveal the differences well. There are many other prosodic aspects (e.g. durations, prosodic voice quality) that remain unchanged and the real differences can be difficult to hear. In [9], better prosodic modeling improved the similarity to the target in a real voice conversion system but the confidence score and the quality score decreased. A sophisticated VC system should retain its quality regardless of whether we are using the conventional MS method or some more advanced approach, and thus we feel that it is best to evaluate the prosody conversion in connection with the spectral conversion.

4.2. Experimental set-up

Experiments in a real voice conversion system were conducted in order to verify the impact of the proposed prosody conversion approach. A general description of the spectral conversion techniques used in the experiments has been given in [6]. The language used in the experiments was English (US). A female voice recorded for TTS purposes served as the source database and several matching sentences were collected from a male speaker. This target speaker was allowed to speak more freely from the prosodic point of view. An interesting observation related to the voices used in the test is that the mean F0 level for the source (female) was 176 Hz and for the target (male) 118 Hz and standard deviations were 18.1 Hz and 15.5 Hz, respectively. However, the mean syllable std in the syllables used in our codebook for the source and the target were 6.7 Hz and 7.1 Hz, respectively. Thus, it is straightforward to see that global std modifications are misleading.

We decided to evaluate the performance of the proposed approach by comparing it against a GMM based pitch prediction model in a listening test. Since the GMM based model and cubic conversion functions were reported to result in quite similar performance as MS [8], the most sophisticated approach of these, i.e. the GMM based technique, was chosen for the experiment. This conventional pitch conversion model was implemented as an 8-mixture GMM.

90 sentences were used for training the spectral conversion and for the training of the proposed prosody conversion approach. 25 sentences, not included in the training set, were used for testing. F0 was measured at every 10 ms and 8 coefficients were used to represent the contour in the transformed domain. The F0 values for the target were generated using two techniques, the GMM based modeling (referred to as GA) and the proposed approach (referred to as CB). The spectral part of the conversion was handled in both cases using identical models and techniques. With the GA method, the durations were not modified as the utterance-level scaling factors were extremely close to 1 for all the test sentences. With the CB method, the durations were modified using the proposed syllable-level scaling. At the syllable level, 22 % of the syllable instances had a scaling ratio falling outside of the range from 0.9 to 1.15.

4.3. Test arrangement

19 listeners participated in the test. Nativeness was not required as the test was designed in such a way that also non-native listeners with good English skills can easily judge the relevant issues from the speech samples. The experiment contained two parts, Test 1 and Test 2. In addition, at the beginning of the test, the subjects were asked to listen to several speech samples from the real target speaker

Table 1. Preference votes given to the proposed approach (CB) and to the GMM based approach (GA), and the ”no preference” votes (equal).

Method	CB	GA	Equal
Test 1	67.0% (318)	22.7% (108)	10.3% (49)
Test 2	70.3% (334)	17.1% (81)	12.6% (60)

(not including the test sentences) and to pay special attention to the speaking style.

In the first part of the test, the listeners heard two versions of the sentences (in which the prosody was converted using the two different techniques, GA and CB). They were asked to choose the sample that best mimicked the target speaker's speaking style. They were guided to choose the sample whose prosody could be closer to the prosody that the target speaker could use. They were asked not to care about quality of the spectral conversion. The subjects could also choose ”equal” and it was possible to listen to the samples as many times as necessary.

The VC system generally leads to somewhat robotic voice quality and the impact that the prosody may have to this phenomenon was studied in the second part of the listening test. The same sentences were played again and the listeners were asked to indicate which sample sounded less robotic. Again, it was possible to respond that the samples were equally robotic.

4.4. Results

The percentages of preference votes that the two methods received as well as the total number of votes are shown in Table 1 for both Test 1 and Test 2. In the first part of the test (Test 1), the results clearly indicate that the proposed approach was found to achieve better prosody conversion than the GMM based approach. In the second part (Test 2), the proposed technique was found to contribute to the voice quality by making it less robotic.

According to a two-tailed t-test, there was a significant difference between the performances of the proposed CB method and the GA method (p=2.9$\cdot 10^{-14}$) for Test 1. Since there was also the third alternative of samples being equally good, the performance of the proposed method was also compared against the summed votes of both the equal choice and the GA method votes. The results were still statistically highly significant (p=9.8$\cdot 10^{-10}$). For Test 2, a similar analysis was performed and the results were also highly significant (p=7.3$\cdot 10^{-19}$ and p=2.3$\cdot 10^{-13}$) for the proposed approach to sound less robotic.

5. CONCLUSIONS

The paper introduced a novel technique for F0 conversion in a voice conversion framework. The proposed approach is based on a prosodic codebook and it aims at finding a compromise between the best possible contour match and the consideration of the linguistic contexts with a limited training data. The F0 contours are represented in transformed domain that allows compression and fast comparison. The proposed method was tested in a real voice conversion system. Listening tests showed very clear preference for the proposed method in mimicking the target speaker's speaking style when compared against the GMM based F0 conversion. The proposed approach was also found to help the overall system in achieving less robotic sound quality.

6. ACKNOWLEDGEMENT

This work has partially been funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech-to-Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org).

7. REFERENCES

[1] Y. Stylianou, O. Cappe, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. on Speech and Audio Processing*, vol. 6(2), pp. 131–142, March 1998.

[2] A. Verma and A. Kumar, “Voice fonts for individuality representation and transformation,” *TSLP*, vol. 2(1), pp. 1–19, February 2005.

[3] O. Turk and L.M. Arslan, “Robust processing techniques for voice conversion,” *Computer Speech and Language*, vol. 4(20), pp. 441–467, October 2006.

[4] M. Abe, S. Nakamura, K.Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” in *ICASSP*, New York, April 1988, pp. 565–568.

[5] J. Yamagishi, K. Ogata, Y. Nakano, J. Isogai, and T. Kobayashi, “HSMM-based model adaptation algorithms for average-voice-based speech synthesis,” in *ICASSP*, Toulouse, May 2006, vol. I, pp. 77–80.

[6] J. Nurminen, V. Popa, J. Tian, Y. Tang, and I. Kiss, “A parametric approach for voice conversion,” in *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, June 2006, pp. 225–229.

[7] D.T. Chapell and J.H. Hansen, “Speaker-specific pitch contour modelling and modification,” in *ICASSP*, Seattle, May 1998, pp. 885–888.

[8] Z. Inanoglu, “Transforming a pitch in a voice conversion framework,” M.S. thesis, University of Cambridge, 2003.

[9] O. Turk and L.M. Arslan, “Voice conversion methods for vocal tract and pitch contour modification,” in *Eurospeech*, Geneve, September 2003, pp. 2845–2848.

[10] B. Gillet and S. King, “Transforming f0 contours,” in *Eurospeech*, Geneve, September 2003, pp. 101–104.

[11] A. Radford, M. Atkinson, D. Britain, H. Clahsen, and A. Spencer, *Linguistics: An Introduction*, Cambridge University Press, Cambridge, 1999.

[12] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, chapter 8, Wiley and Sons, New York, 2001.

[13] T. Ceyssens, W. Verhelst, and P. Wambacq, “On the construction of a pitch conversion system,” in *EUSIPCO*, Toulouse, September 2002.