LANGUAGE-INDEPENDENT ACOUSTIC CLONING OF HTS VOICES: A PRELIMINARY STUDY

Carmen Magariños[1], *Daniel Erro*[2], *Eduardo R. Banga*[1]

[1]Multimedia Technology Group (GTM), AtlantTIC Research Centre, University of Vigo, Spain
[2]IKERBASQUE – Aholab, University of the Basque Country, Bilbao, Spain

ABSTRACT

This paper presents a new method for cross-lingual speaker adaptation in the framework of HMM-based speech synthesis. Taking two HTS voice models as input, one for the desired language and another for the aimed speaker identity, it yields a third model that produces speech in the target language while sounding like the target speaker. The method operates at segmental level (spectral information and average fundamental frequency) and does not require any phonetic or linguistic information. Perceptual evaluation experiments show that, when the input models are good enough, the resulting synthetic voice is perceived as similar to the target speaker with no important quality degradation.

Index Terms— HMM-based speech synthesis, Cross-lingual speaker adaptation, Polyglot synthesis, Multilingual synthesis

1. INTRODUCTION

Text-to-Speech (TTS) systems have traditionally required a costly process of recording new voices in order to synthesize new speakers, speaking styles, emotions or languages. However, the emergence of statistical parametric speech synthesis [1] made possible the modification of the speech characteristics and/or speaker identity in a quite flexible way, avoiding the need of long additional recordings. In particular, HMM-based speech synthesis is able to make such modifications by means of speaker adaptation techniques [2, 3]. Nevertheless, up to now, most of the work in this field has been focused on *intra-lingual* speaker adaptation (source and target speakers speak the same language) while the *cross-lingual* paradigm (source and target speakers speak different languages) seems to be less explored.

Regarding *cross-lingual* adaptation, in [4] the authors propose a HMM-based method for synthesizing speech in multiple languages from a single language-independent acoustic model. This acoustic model is first trained from speech data of several speakers in different languages and then it may be adapted to any specific speaker to obtain a speaker dependent (SD) model. As a result, it is possible to synthesize speech from any SD model in any language in the training set. In [5] and [6] a mobile device that implements personalized speech-to-speech translation is presented. Among the integrated algorithms in this device, there is a cross-lingual speaker adaptation method which is based on a state-level mapping, first proposed in [7]. This mapping makes use of the minimum Kullback-Leibler divergence (KLD) between paired HMM states in the input and the output languages. Another method for obtaining polyglot speech synthesis is described in [8]. In this case, the speech is factorized into speaker-specific and language-specific characteristics, which are modeled by separate transforms. Thus, language and speaker features can be independently controlled.

Within the standard HMM-based synthesis framework [9], this paper proposes a new method to combine the language-dependent structure of a synthesis voice model with the acoustic characteristics of another model trained for a different language. Unlike classical adaptation techniques, where the source model is transformed to fit some input data from a target speaker, the proposed technique transforms the source model to be acoustically closer to another model that conveys the identity of the target speaker. As a result, a third model is built, which is able to produce synthetic speech in the same language as the source model using the target speaker's voice. The proposed method uses the INCA algorithm [10, 11] to align the states of the two involved models; then, a transformation function is trained to transform the acoustic emission distributions of the source model states into those of the target speaker's model. The main advantage of the method is its language independent condition: model-to-model adaptation is performed rapidly without any phonetic or linguistic information. In exchange, it can deal only with the segmental characteristics of voice, i.e. spectral information (more specifically, Mel-cesptral representation of the spectral envelope) and mean fundamental frequency.

The proposed method is applicable, for instance, to build a multilingual speech synthesizer with a unique voice. Using heterogeneous models already available for each of the involved languages, this can be done by cloning one of the speakers in all of these languages. In the context of person-

This work has been supported by the Galician Regional Government (CN2011/019, CN2012/160), the European Regional Development Fund and the Spanish Government (BES-2013-063708 and TEC2012-38939-C03).

alized speech-to-speech translation, a voice model of the user in his/her own language can be obtained by means of standard adaptation techniques; then, the speaker identity can be transferred to a model in the desired language thanks to the proposed method. Alternatively, the method can be applied to homogenize the average voice models of several languages, so that adaptation transforms trained for one language can be applied to all of them.

The remainder of this paper is structured as follows: section 2 explains the proposed adaptation method; section 3 presents the evaluation experiments and a discussion of the results; and finally, section 4 summarizes the main conclusions.

2. DESCRIPTION OF THE METHOD

Given two HTS voice models in two different languages, let's say model 1 and model 2, the goal is to obtain a new model, model 3, with the linguistic structure of model 1 and the acoustic properties of the voice given by model 2. The idea is illustrated in Fig. 1. The proposed method estimates a mapping between the emission distributions (Gaussian p.d.f.'s) of both models, and then projects the distributions of model 1 onto those of model 2 without altering its linguistic structure. For clarity, we will refer to model 1 as *source speaker model* and model 2 as *target speaker model*.

Each of the input models is composed by three different acoustic streams: logarithm of the fundamental frequency ($\log F0$), Mel-cepstral representation of the spectral envelope, and excitation parameters. This preliminary study is focused mainly on transforming the cepstral part of the source model without using any linguistic feature. As for $\log F0$, since the specific intonation patterns of each language (and also each speaker) are difficult to capture with a set of relevant acoustic features, we apply a simple mean normalization. The excitation, the durations and the global variance statistics of the source voice model are kept unmodified.

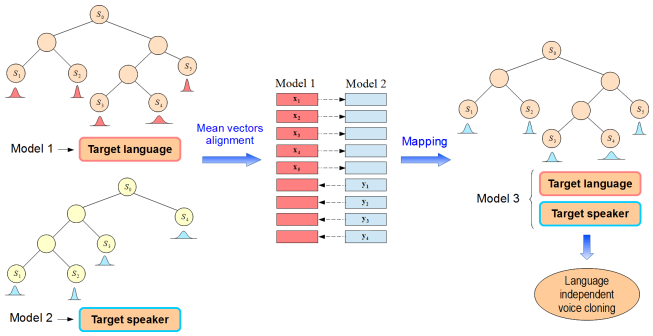

Fig. 1. Cloning of HTS voices.

2.1. MCEP adaptation

The cepstral information of the source model is projected onto the cepstral space of the target speaker's model. We can distinguish two stages: one initial stage for the alignment of the distributions of both models, and a second one for estimating the final mapping between them. Fig. 2 shows the block diagram of the whole process.

2.1.1. Alignment of mean vectors

In the standard HTS configuration, five emitting states per phone are considered; thus, states are arranged in five different decision trees. The first step of this stage is to compact, for both models, the whole set of Mel-cepstral distributions of the five trees. In practice, we found it was convenient to consider for alignment only the static part of the mean vectors of the distributions. Also, in order to take into account the relative importance of the distributions within the model, each one is assigned a weight that is proportional to the occupancy of that state when the model was trained. From here on, we will refer to the static part of the source mean vectors as $X = \{\mathbf{x}_i\}_{i=1\ldots N_x}$ and to the target one as $Y = \{\mathbf{y}_j\}_{j=1\ldots N_y}$. Their weights will be referred to as $\{w_i\}$ and $\{v_j\}$ respectively. We will assume that these weights are normalized: $\sum_{i=1}^{N_x} w_i = \sum_{j=1}^{N_y} v_j = 1$.

The next step is to find an acoustic correspondence between the states of the two models. To do that, $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_j\}$ are paired using a modified version of the INCA algorithm [10], which was originally proposed for a similar purpose in the voice conversion field. Our implementation of INCA consists of the following steps:

1. Initialization of an auxiliary vector set X': $X' = X$
2. Bi-directional nearest neighbor search between the vectors in X' and those in Y, allowing repetitions:

$$\hat{i} = \underset{j=1\ldots N_y}{\arg\min} \|\mathbf{x}'_i - \mathbf{y}_j\| \;, \quad \hat{j} = \underset{i=1\ldots N_x}{\arg\min} \|\mathbf{y}_j - \mathbf{x}'_i\| \quad (1)$$

3. Training of a linear projection function $F(\mathbf{x}) = \mathbf{A}\mathbf{x}+\mathbf{b}$ between X and Y, given the current index pairs $\{i \leftrightarrow \hat{i}\}$ and $\{\hat{j} \leftrightarrow j\}$:

$$\begin{aligned}\{\mathbf{A}, \mathbf{b}\} = \arg\min \; & \textstyle\sum_{i=1}^{N_x} p_{i,\hat{i}} \|F(\mathbf{x}_i) - \mathbf{y}_{\hat{i}}\|^2 + \\ & + \textstyle\sum_{j=1}^{N_y} p_{\hat{j},j} \|F(\mathbf{x}_{\hat{j}}) - \mathbf{y}_j\|^2\end{aligned} \quad (2)$$

Unlike the original implementation of INCA, the weight of each pair is calculated as

$$p_{i,j} = w_i \frac{v_j}{V_i} + v_j \frac{w_i}{W_j} \quad (3)$$

where V_i is the sum of the individual weights of all the vectors paired with $\mathbf{x}'_i$ and W_j is the sum of the weights of the vectors paired with $\mathbf{y}_j$, with possible repetitions.

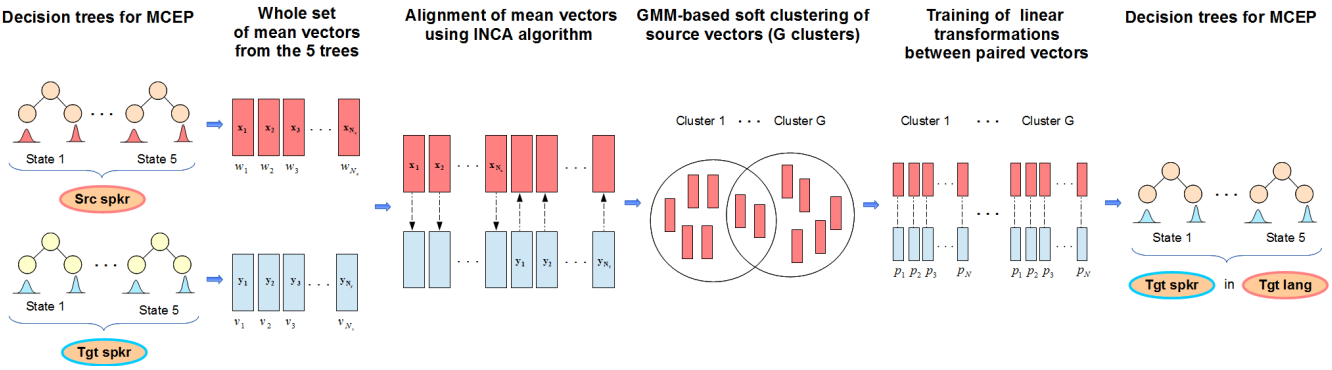

Fig. 2. Block diagram of the MCEP adaptation.

4. Update of X' in accordance with the new F: $X' = \{\mathbf{x}'_i\}, \mathbf{x}'_i = F(\mathbf{x}_i)$
5. If the maximum number of iterations has been reached or there are no changes with respect to the last iteration, exit; otherwise, go back to step 2.

In the general case, the minimization at step 3 can be carried out by solving an overdetermined set of equations. For clarity, let us now denote the N source-target pairs at a given INCA iteration as $\{\mathbf{x}_n, \mathbf{y}_n\}$, with weights $\{p_n\}$. Under this notation, the solution of eq. (2) is

$$\mathbf{W} = [\mathbf{A}\ \mathbf{b}]^\top = (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}^\top \mathbf{Y} \tag{4}$$

where

$$\hat{\mathbf{X}} = \begin{bmatrix} p_1 \hat{\mathbf{x}}_1^\top \\ \vdots \\ p_N \hat{\mathbf{x}}_N^\top \end{bmatrix}, \quad \hat{\mathbf{x}}_n^\top = [\mathbf{x}_n^\top\ 1], \quad \mathbf{Y} = \begin{bmatrix} p_1 \mathbf{y}_1^\top \\ \vdots \\ p_N \mathbf{y}_N^\top \end{bmatrix} \tag{5}$$

In practice, however, N is low in comparison with the number of unknowns of $\mathbf{W}$. To overcome this problem, $\mathbf{A}$ is forced to be a band-matrix with an adjustable number of diagonal bands. In that case, the non-zero elements of $\mathbf{W}$ are computed in a column-by-column fashion (the mathematical details are omitted because of space limitations).

2.1.2. *Final mapping*

Finally, a transformation function is trained using the definitive set of paired vectors. Before that, to compensate for one-to-many alignment, for each source vector paired with more than one target vector, we keep only the target vector with the largest weight. At the same time, duplicate pairs are eliminated. As a result, we get a final set of N_x vector pairs denoted $\{\mathbf{x}_n, \mathbf{y}_n\}$. Since some of the initial target vectors may not take part in the final set, the weights $\{v_n\}$ are renormalized so that their sum equals 1 again.

Before training the final transformation, we perform a soft classification of the source vectors $\{\mathbf{x}_n\}$ using a Gaussian mixture model (GMM) of G components, denoted Θ. Then, a probabilistic combination of linear transforms given by

$$F(\mathbf{x}) = \sum_{g=1}^{G} P(g/\mathbf{x}, \Theta)[\mathbf{A}_g \mathbf{x} + \mathbf{b}_g] \tag{6}$$

is trained via error minimization:

$$\{\mathbf{A}_g, \mathbf{b}_g\}_{g=1\ldots G} = \operatorname{argmin} \sum_{n=1}^{N} p_n \|\mathbf{y}_n - F(\mathbf{x}_n)\|^2 \tag{7}$$

where $p_n = w_n + v_n$. Similar transforms are trained also for the dynamic parts of the mean vectors of the source and target distributions, $\{\Delta\mathbf{x}_n, \Delta\mathbf{y}_n\}$ and $\{\Delta^2\mathbf{x}_n, \Delta^2\mathbf{y}_n\}$, which had been kept aside until now. The training procedure is similar to that in [12], with some modifications to control the number of diagonal bands of the involved matrices. Finally, to get the output voice model, the mean Mel-cepstral vectors of all the distributions in the source model are replaced by their transformed counterparts.

2.2. F0 adaptation

Similarly as in the Mel-cepstral part, the first step is to compact the distributions of the five decision trees created for $\log F0$. Once again, we only use the static part of the mean vectors, denoted as $\{x_i\}$ and $\{y_j\}$ (in this case these are scalars), to perform the adaptation. Since the $F0$ is modelled through Multi-space distributions (MSD) [13], the weight assigned to each distribution is calculated as the product of the normalized state occupancies and its so-called MSD weight (which can be interpreted as the probability of voicing). The resulting source and target weights are referred to as $\{\check{w}_i\}$ and $\{\check{v}_j\}$ respectively. Given their strong linguistic dependencies, intonation patterns are beyond the scope of this work; instead, we perform a simple average $\log F0$ normalization expressed as

$$F(x_i) = x_i - \frac{\sum_{i=1}^{N_x} \check{w}_i x_i}{\sum_{i=1}^{N_x} \check{w}_i} + \frac{\sum_{j=1}^{N_y} \check{v}_j y_j}{\sum_{j=1}^{N_y} \check{v}_j} \tag{8}$$

3. EXPERIMENTS AND DISCUSSION

Before the formal evaluation of the system, its configuration was optimized by means of informal listening tests. The best performance was achieved for the following settings:

- The 0^{th} Mel-cepstral coefficient is not considered, neither for the alignment nor for the final mapping of mean vectors. States corresponding to silences are also excluded.
- Only the first 15 static Mel-cepstral coefficients are used for the alignment. Using more coefficients increases the computational load with no audible improvement.
- Band-matrices of radius 4 (9 diagonal bands) are used for both alignment and transformation.
- The maximum number of INCA iterations is set to 25.

The proposed method was evaluated by means of a perceptual listening test. Since, to the best of the authors' knowledge, there was no baseline method that could be used as reference for comparison, we calculated differential mean opinion score (DMOS) for two aspects of the cloned voices: similarity to the target speaker, and quality in comparison with the source synthetic voice.

We used speech databases in four different languages: English, Castilian Spanish, Basque and Galician. We selected two female and two male synthetic voice models: FS (Spanish female), FG (Galician female), MB (Basque male) and ME (English male). The amount of training utterances per voice was approximately 2000, 10000, 4000 and 2800, respectively. The models were trained using version 2.2 of HTS. Acoustic analyses were performed using the same vocoder, namely Ahocoder [14], and language-specific text analyzers were used for each voice [15, 16, 17]. To evaluate the performance of the system under all possible gender combinations, we chose the following conversion directions: FS-MB, FG-FS, ME-FG and MB-ME, where the first element of each pair represents the source speaker (which provides the language of the output voice) and the second one the target speaker (which provides the speaker identity).

A total of 23 listeners participated in the test, 10 of which were familiar with speech processing techniques. All of them were fluent speakers of at least three of the languages of the evaluation, namely Spanish, English and either Galician or Basque, and were slightly familiar with the fourth one. Each listener was presented, in random order, with 12 randomly selected trios of samples (3 per conversion direction): source speaker, target speaker and converted speaker. For each trio, listeners were asked to give two opinion scores on a 5-point scale: one score to punctuate how similar the third sample was to the second one, and a second score to rank the quality of the third sample in comparison with the first one. The evaluation was conducted through a web interface, and listeners were asked to use headphones.

Fig. 3 shows the similarity and quality DMOSs obtained for each conversion direction, along with the corresponding 95% confidence intervals. Global average scores are provided too. Remarkably, there are two pairs of voices for which the results are particularly good (nearly 4 points in both performance dimensions), while for the other two combinations the performance was not that good (3 points or lower). The reason for this dichotomy seems to be the quality of the synthetic voice used as source in each case. Indeed, the quality of the original female voices was higher than that of the male voices due either to the careful manual segmentation of the database (FS) or to the amount of training data (FG). This is a valuable observation which, nevertheless, needs to be confirmed by more extensive experiments. Regarding the pair with the lowest scores, namely MB-ME, it is worth mentioning that ME exhibited much larger $\log F0$ variance than MB, which probably contributed to the observed scores. It is also worth considering that the way the output voice was shown to the evaluators, i.e. preceded not only by a sample of the target speaker's voice but also by the same utterance in the source speaker's voice, may have produced an overall decrement of the similarity scores, as the linguistic and suprasegmental characteristics of the source voice are not modified. As a general conclusion, we can state that there are at least some voices, presumably high-quality voices, not necessarily from the same gender, for which the proposed method gives a satisfactory performance.

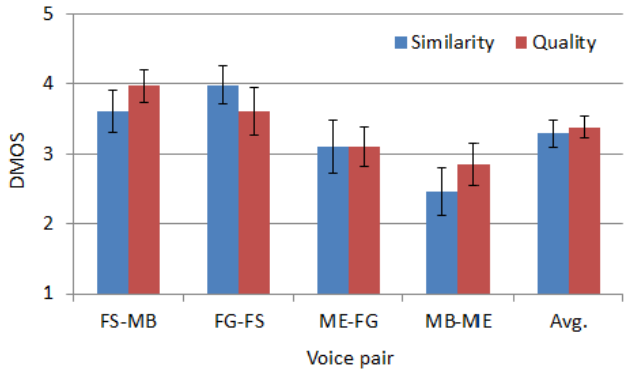

Fig. 3. DMOS results (examples at http://goo.gl/FwemL4).

4. CONCLUSION

In this paper we have described a new method for cross-lingual speaker adaptation. From two (source and target) HTS voice models in different languages, the proposed algorithm builds a third model that combines the language of the source model with the speaker identity of the target model. Since the method works on a purely segmental level, it does not need any language-specific information. Perceptual tests with different pairs of speakers (and languages) have shown the potential of the method as long as the starting models are good enough.

5. REFERENCES

[1] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, November 2009.

[2] J. Yamagishi, *Average-voice-based speech synthesis*, Ph.d. dissertation, Tokyo Inst. of Technol., Yokohama, Japan, 2006.

[3] J. Yamagishi, Z.-H. Ling T. Toda T. Nose, H. Zen, K. Tokuda, S. King, and S. Renals, "Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 6, pp. 1208–1230, August 2009.

[4] J. Latorre, K. Iwano, and S. Furui, "New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer," *Speech Communication*, vol. 48, pp. 1227–1242, May 2006.

[5] K. Oura, J. Yamagishi, M. Wester, S. King, and K. Tokuda, "Analysis of unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using KLD-based transform mapping," *Speech Communication*, vol. 54, pp. 703–714, January 2012.

[6] J. Dines, H. Liang, L. Saheer, M. Gibson, W. Byrne, K. Oura, K. Tokuda, J. Yamagishi, S. King, M. Wester, T. Hirsimki, R. Karhila, and M. Kurimo, "Personalising speech-to-speech translation: Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis," *Computer Speech and Language*, vol. 27, pp. 420–437, September 2013.

[7] Y. Nankaku Y. J. Wu and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Proc. of Interspeech*, 2009, pp. 528–531.

[8] H. Zen, N. Braunschweiler, S. Buchholz, M. Gales, K. Knill, S. Krstulovic, and J. Latorre, "Statistical Parametric Speech Synthesis Based on Speaker and Language Factorization," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 6, pp. 1713–1724, August 2012.

[9] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. 6th ISCA Speech Synthesis Workshop*. ISCA, 2007, pp. 294–299.

[10] D. Erro, A. Moreno, and A. Bonafonte, "INCA Algorithm for Training Voice Conversion Systems From Nonparallel Corpora," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 5, pp. 944–953, July 2010.

[11] Y. Agiomyrgiannakis, "The Matching-Minimization algorithm, the INCA algorithm and a mathematical framework for voice conversion with unaligned corpora," in *Proc. of ICASSP*, 2016, pp. xxx–xxx.

[12] H. Ye and S. Young, "Perceptually Weighted Linear Transformations for Voice Conversion," in *Proc. Eurospeech: 8th European Conference on Speech Communication and Technology*, 2003.

[13] K. Tokuda, T. Mausko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D, no. 3, pp. 455–464, March 2002.

[14] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis," *IEEE Journal Sel. Topics in Signal Process.*, vol. 8, no. 2, pp. 184–194, April 2014.

[15] E. Rodríguez-Banga, C. García-Mateo, F. Méndez-Pazó, M. González-González, and C. Magariños, "Cotovía: an open source TTS for Galician and Spanish," in *Proc. IberSPEECH 2012: VII Jornadas en Tecnologa del Habla and III Iberian SLTech Workshop*. RTTH and SIG-IL, 2012, pp. 308–315.

[16] I. Sainz, D. Erro, E. Navas, I. Hernáez, J. Sánchez, I. Saratxaga, I. Odriozola, and I. Luengo, "Aholab speech synthesizers for albayzin2010," in *Proc. FALA'2010*, 2010, pp. 343–348.

[17] P. Taylor, A. W. Black, and R. Caley, "The architecture of the Festival speech synthesis system," in *Proc. of the ESCA Workshop in Speech Synthesis*, 1998, pp. 141–151.