Speaker Adaptation using only Vocalic Segments via Frequency Warping

Agustin Alonso¹, Daniel Erro¹,², Eva Navas¹, Inma Hernaez¹

¹AHOLAB, University of the Basque Country (UPV/EHU), Bilbao, Spain
²Basque Foundation for Science (IKERBASQUE), Bilbao, Spain

agustin@aholab.ehu.es, derro@aholab.ehu.es, eva@aholab.ehu.es, inma@aholab.ehu.es

Abstract

Speaker adaptation techniques allow hidden Markov model (HMM) based speech synthesis systems to mimic a target voice of which a few samples are available. However, usual adaptation approaches are not applicable when the target voice is dysarthric, i.e. the target speaker has an impairment which prevents the correct pronunciation of some phonemes. As a first step towards giving personalized synthetic voices to these particular speakers, this paper explores the possibility of adapting the whole statistical voice model using frequency warping (FW) based transformations trained exclusively with vowels. Perceptual evaluations performed for healthy voices show that the proposed method achieves reasonable results even when the adaptation data exhibit medium/low recording quality.

Index Terms: speech adaptation, statistical speech synthesis, frequency warping, dysarthric voice

1. Introduction

The voice is one of the most natural methods we have to communicate. Sadly, some people have lost partially or totally their voice either due to an accident or to some illness. Speech technologies can alleviate this situation, and in particular a Text-to-Speech (TTS) system [1] can be used by people with speech impairments to improve their communication. The recent development of smartphones, together with the fact that statistical parametric TTS systems require nowadays few resources [2], have made possible that anybody can have his/her own TTS in the pocket. Although this has made easier for these people to communicate in a natural way, conventional TTS systems produce a generic and personality-less voice. An important step towards users’ acceptance of TTS for daily communication is the ability to give a personalized and unique voice to each user. Hence, there are various initiatives to provide personalized voices to these users using voice adaptation [3]; with a small amount of recordings of a human target voice, it is possible to change the acoustic properties of the synthetic voice.

Adaptation is feasible when a user knows when he/she will lose his/her voice, for example in the case of a planned surgery. Then, he/she can make the necessary recordings before that. However, this is not an option when the user’s voice exhibits severe symptoms of a given pathology. In this situation, there are two possible approaches. The first one is to use recordings made by “donors”. In other words, the user can borrow the acoustic properties of a donor’s voice [4]. This is certainly the most reasonable strategy when the user has completely lost the ability of speaking. The second approach makes sense when the user is able to produce speech but, due to impairment, only some sounds are correctly pronounced, i.e. a dysarthric voice.

Then, it is possible to extract characteristics from the speech segments correctly produced and try to extrapolate this information to adapt the whole voice [5].

Following this second approach, in this paper we present a method to adapt statistical parametric synthetic voices using speech segments instead of whole utterances. More specifically, we suggest performing spectral adaptation using only vocalic fragments to train the necessary transforms, as vowels are in general the most reliable speech segments from a dysarthric voice. Instead of using unconstrained linear transforms as usual adaptation methods do, our method is based on constrained linear transforms with explicit physical meaning [6, 7], which are easier to extrapolate to phonemes other than vowels. Moreover, these transforms are known to preserve well the quality of the voice, which can be of primary interest for speaking impaired. In exchange, the capability of accurately mimicking target voices is slightly reduced. In this first stage of the work, we have tested the method using healthy voices. This allows us to develop and evaluate the different setups using well-established methods as baseline, thus getting an indicator of how far we can get following this strategy.

The rest of the paper is structured as follows. Section 2 contains a brief description of the background technology. In section 3 the proposed method and its implementation are explained in detail. In section 4 the experiments carried out to prove the method and its results are presented. Finally in section 5 the conclusions of this work are summarised.

2. General background

2.1. Statistical parametric speech synthesis

Statistical parametric synthesis has gained popularity in the last ten years due to its several advantages over alternative technologies [2]: smooth synthetic output without discontinuities, unsupervised training, need of relatively few training data, low footprint, and flexibility to change some aspects of the voice such as emotional expression or speaker identity. Another important reason for the rapid growth of statistical synthesis is the availability of HTS (HMM-based Speech Synthesis System) [8], an open source system which allows easily training, using and adapting synthetic voices.

During the training phase, a database containing audio utterances from one (or more) speaker(s) and their transcriptions is used to train a set of HMMs. Texts are translated into sets of phonetic, prosodic and linguistic labels using a text analyzer, and a vocoder is used to extract acoustic feature vectors from speech waveforms. State-of-the-art vocoders [9, 10] consider three different acoustic streams: logarithm of the fundamental frequency ($f_0$), Mel-cepstral (MCEP) or linear prediction related representation of the spectral envelope, and degree of harmonicity of different spectral bands. The context-dependent
HMMs learned by the system capture the correspondence between the labels and the acoustic parameters, together with their first and second derivatives over time. For a more adequate modeling of durations, a variant of HMM called hidden semi-Markov model (HSMM) [11] is used in which state durations are characterized by explicit Gaussian distributions instead of being tied to transition probabilities.

In the synthesis phase, any arbitrary input text is translated into labels by the same text analyser used during training; given these labels and the model, the system generates the most likely sequence of acoustic parameters [12] and feeds it into the vocoder, which builds the corresponding synthetic waveform.

The mean vectors and covariance matrices of the Gaussian emission distributions of the HSMM states can be modified to change the acoustic properties of the voice. Indeed, it is possible to adapt them to a small amount of recordings from a new speaker — the target speaker — using algorithms such as Constrained Maximum Likelihood Linear Regression (CMLLR) [13] or Constrained Structural Maximum-a-Posteriori Linear Regression (CSMAPLR) [14]. In general, the source voice model of speaker-adaptive systems is an average voice model, i.e. a model trained from a database containing several different speakers [15]. Average voice models are especially suitable for adaptation because they do not convey peculiarities from any particular speaker, so they are easier to adapt to new data.

2.2. Frequency warping plus amplitude scaling

Usual adaptation techniques (CMLLR, CSMAPLR, etc.) use linear transforms to project the distributions of the source voice model onto the target speaker’s acoustic space. Unconstrained linear transforms are flexible enough to capture the source-target correspondence but are not directly interpretable. On the contrary, there is a different type of transform, namely Frequency Warping (FW) plus Amplitude Scaling (AS), which is not as flexible as the linear one but has a clear physical interpretation [7]. FW maps the frequency axis of a source spectrum onto that of a target spectrum, without eliminating any spectral detail from the source. Hence, this operation does not degrade the quality of speech significantly. AS compensates for the amplitude differences between the frequency-warped source spectrum and the target spectrum. FW+AS transforms were originally used in the voice conversion field [6, 16, 17, 18], where they led to satisfactory conversion scores as well as high quality scores. Remarkably, it has been shown that, in the cepstral domain, FW is equivalent to a multiplicative matrix [19] and AS can be implemented by means of an additive cepstral term. In other words, FW+AS is linear in the cepstral domain. Therefore, it can be applied to adapt HSMM state distributions, as generic linear transforms do:

\[
\bar{\mu} = A\mu + \bar{b}, \quad \bar{\Sigma} = A\Sigma A^T
\]

where \(\mu\) and \(\Sigma\) are the mean vector and covariance matrix of the HSMM state, \(\bar{\mu}\) and \(\bar{\Sigma}\) are their transformed counterparts, and

\[
a = \begin{bmatrix} A & 0 & 0 \\ 0 & A & 0 \\ 0 & 0 & A \end{bmatrix}, \quad b = \begin{bmatrix} b \\ 0 \\ 0 \end{bmatrix}
\]

In this case \(A\) represents the FW function and \(b\) the AS. The block structure of \(A\) and \(b\) allows to transform directly the static and the dynamic features of the acoustic model. In the context of this work, the advantage of this transformation lies in the fact that, given its physical constraints, it can be carefully extrapolated to phonemes not seen during training without producing uncontrolled distortions. Next section shows how the proposed method exploits this key property.

3. Proposed method

3.1. Selection of training data

Though the long-term goal of this work is performing adaptation using pathological voices, in this primary stage we have used healthy voices to develop the main algorithms. The motivation for this is simplifying the problem and getting a first indicator of the performance we can expect from the proposed method in the most favorable conditions. Moreover, comparison with state-of-the-art adaptation methods is possible only using healthy voices. Thus, we used labelled speech material to facilitate the selection of the vocanic segments.

The FW+AS-based adaptation methods described below require a set of paired source and target vectors as input. The target frames are taken from the vowels uttered by the target speaker. More specifically, we take the central frame of the vowels where: (i) all the frames are voiced, which avoids artifacts related to \(f_0\) misdetection; (ii) the duration is greater than 55ms, which ensures that there are enough samples for an accurate spectral analysis and also that there is a sufficiently wide stable zone free of co-articulation. The source frames are built from the original voice model: first, we extract labels from the text of the adaptation utterance; using these labels, we select the p.d.f. of the central state of each vowel (in a standard five-state HSMM configuration, the third one), which represents the most stable part of the phoneme; finally, we take the static part of its mean vector as source frame.

3.2. Dynamic frequency warping

To learn the FW transformation from the adaptation data, we use a modified version of the method proposed in [6], which is based on the Dynamic Frequency Warping (DFW) procedure [20]. DFW calculates the FW function that should be applied to a set of \((N + 1)\)-point log-amplitude semispectra, \(\{X_i\}\), to make them maximally close to their paired counterparts, \(\{Y_i\}\). It is based on a cost function \(D(i, j)\) that indicates the accumulated log-spectral distortion obtained when the \(i^{th}\) bin is mapped into the \(j^{th}\) bin of the target spectra following the “best” path from \((0, 0)\) to \((i, j)\). \(D(i, j)\) can be expressed mathematically as follows:

\[
D(i, j) = \min \left\{ D(i-1, j) + d(i, j), D(i-1, j-1) + w \cdot d(i, j) \right\}
\]

(3)

where \(i, j = 0 \ldots N, w\) is a parameter that controls the penalty of the vertical and horizontal paths, and \(d(i, j)\) is the distance between the \(i^{th}\) bin from the source and the \(j^{th}\) bin from the target. In our case, \(d(i, j)\) is calculated simultaneously from all the available pairs of training vector using the following equation:

\[
d(i, j) = \sum_{t=1}^{T} (X_i[t] - Y_i[j])^2 + \sum_{t=1}^{T} \alpha_i (X_i'[t] - Y_i'[j])^2
\]

(4)

where \(T\) is the total amount of training pairs, \(X_i\) and \(Y_i\) are the derivatives of \(X_i\) and \(Y_i\) over frequency, and \(\alpha\) is an empirical factor that stands for the relative weight of the derivatives of the spectra. The first part of (4) captures the difference of absolute local spectral amplitudes, and the second one helps align spectral events given by abrupt slopes, i.e. formants - this second
term was not considered in [6]. The frequency warping path \( P \) is defined as a sequence of points,

\[
P = \{(0, 0), (i_1, j_1), (i_2, j_2), \ldots (N, N)\}
\]

(5)
such that the presence of \((i, j)\) in \( P \) indicates that the \( j^{th} \) bin of the source spectrum should be mapped onto the \( j^{th} \) bin of the target spectrum. The points of \( P \) are backtracked from \((N, N)\) following the minimal-distortion path in inverse order using the recursion expressed in (3).

Since DFW works in the log-spectral domain, the paired source and target \( p^{th} \)-order MCEP vectors have to be converted into \((N + 1)\)-point discrete log-amplitude semispectra. Using a traditional MCEP definition, this can be done multiplying the MCEP vectors by a matrix \( S \) whose elements are defined as follows

\[
S[n, i] = \cos(i \cdot \text{mel}(\pi n/N)), \quad 0 \leq n \leq N, \quad 0 \leq i \leq p
\]

(6)

Similarly, the \( p^{th} \) order representation of a discrete log-amplitude spectrum can be calculated through the technique known as regularized discrete cepstrum [21], i.e. multiplying the \((N + 1)\) point discrete log-amplitude semispectrum in vector form by

\[
C = \left(S^T S + \lambda R\right)^{-1} S^T \rho
\]

(7)

where \( S \) is given by (6), \( R \) is a regularization matrix that imposes smoothing constraints to the cepstral envelope,

\[
R = 8\pi^2 \cdot \text{diag}\{0, 1^2, 2^2, \ldots, p^2\}
\]

(8)

and \( \lambda \) is an empirical constant typically equal to \(2 \cdot 10^{-14}\) [21]. In practice, the \( 0^\text{th} \) cepstral coefficient, related to energy, and the \( 1^\text{st} \) cepstral coefficient, mainly related to glottal spectrum, are not relevant in terms of FW [6], so they are set to zero before the multiplication by \( S \). Once the training MCEP vectors are translated into log-spectral domain using \( S \), an optimal warping path \( P \) is obtained via DFW. As detailed in [6], the matrix that implements the frequency warping operation in the log-spectral domain can be defined as

\[
W[j, i] = \frac{m_{i,j}}{\sum_{k=1}^{N} m_{i,k}}
\]

(9)

where \( m_{i,j} = 1 \) when \((i, j) \in P \) and \( m_{i,j} = 0 \) otherwise. Note that the denominator of (9) compensates for the one-to-many mappings between source and target bins, which are inevitable according to the recursion in (3) and the resulting structure of \( P \). Once \( W \) has been determined, it is possible to calculate the matrix \( A \) that implements the same operation in the MCEP domain:

\[
A = C \cdot W \cdot S
\]

(10)

In this work, we calculate a single FW matrix \( A \) for the whole training dataset, regardless of the vowel the vectors belong to. Using \( A \) and its block-replicated version \( \tilde{A} \) (2), this curve can be applied to any distribution of the model without altering its phonetic content.

3.3. Amplitude scaling

The additive MCEP term that implements the AS operation is calculated as the difference between warped-source and target MCEP vectors:

\[
b = \frac{1}{T} \sum_{t=1}^{T} (y_t - A \cdot x_t)
\]

(11)

where \( x_t \) and \( y_t \) are the source and target MCEP vectors in pair \( t \), respectively, and \( A \) is given by (10). Instead of assuming a single AS vector, \( b \) is trained separately for each vowel. Unlike FW, when performing the model adaptation via expressions (1) and (2), each distribution in the model is given an individual AS vector. Three possible cases are considered:

- Distributions that correspond to only one vowel are modified by the specific AS vector of that vowel.
- When a distribution is shared by more than one vowel, we use a weighted average of the AS vectors of these vowels. The weights are the proportion of times the current distribution belonged to each vowel when the original voice model was trained.
- The remaining distributions are modified using the average of the AS vectors of all vowels.

3.4. Average fundamental frequency correction

Until now we have adapted only the spectral information. Another very important acoustic feature that defines the identity of the speaker is \( f_0 \). To adapt it, we perform a single mean normalization as follows. First, we take the log-\( f_0 \) values at the central frame of the vowels. From these values, we calculate the target average log-\( f_0 \). Then, similarly as in the spectral case, a source average log-\( f_0 \) is calculated from the static parts of the means of the p.d.f. in the central state of the vowels. Adaptation is carried out by applying a linear transformation similar to (1) to all the distributions in the log-\( f_0 \) model, for \( A = 1 \) and \( b \) equal to the difference between source and target average log-\( f_0 \).

4. Experiments

4.1. Experimental setup

We have tested the proposed method using recordings made through a voice-banking dedicated web-portal [22]. This type of data is the most adequate for our experiments because, in the near future, we plan to integrate adaptation functionalities in the website in order to provide speaking aids to impaired people. We collected 100 phonetically balanced sentences from 6 non-professional Spanish speakers (3 males, 3 females). All of them were recorded using the speaker’s own equipment and over a variable number of sessions. As a result, the recordings exhibit medium/low quality. All signals were normalized in order to maximize the dynamic range, and were passed through a Wiener filter to reduce the noise. Phonetic segmentation was carried out automatically using HTK [23]. Table 1 shows the number of vocalic segments obtained according to the process described in subsection 3.1. As expected, each speaker has a different amount of usable vocalic segments.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>/a/</th>
<th>/e/</th>
<th>/i/</th>
<th>/o/</th>
<th>/u/</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>256</td>
<td>154</td>
<td>111</td>
<td>153</td>
<td>66</td>
<td>740</td>
</tr>
<tr>
<td>M2</td>
<td>451</td>
<td>239</td>
<td>182</td>
<td>249</td>
<td>63</td>
<td>1184</td>
</tr>
<tr>
<td>M3</td>
<td>281</td>
<td>218</td>
<td>141</td>
<td>150</td>
<td>64</td>
<td>947</td>
</tr>
<tr>
<td>F1</td>
<td>317</td>
<td>201</td>
<td>169</td>
<td>186</td>
<td>74</td>
<td>1530</td>
</tr>
<tr>
<td>F2</td>
<td>453</td>
<td>364</td>
<td>244</td>
<td>334</td>
<td>135</td>
<td>1123</td>
</tr>
<tr>
<td>F3</td>
<td>368</td>
<td>249</td>
<td>196</td>
<td>226</td>
<td>84</td>
<td>854</td>
</tr>
</tbody>
</table>

As source voice, we used a speaker-dependent HTS voice trained with 1962 utterances from a female speaker in Spanish - the use of an average voice model was discarded because
we did not have access to a sufficiently large amount of speech material of the desired quality. All the audio signals, both for the source voice and for the target voices, were sampled at 16 kHz and parameterized using Ahocoder [10], a high quality harmonic-plus-noise model based vocoder. Ahocoder extracts 39th-order MCEP coefficients, log-f0 and maximum voiced frequency (MVF) each 5 ms.

To evaluate our method we compared it with a state-of-the-art HTS-based speaker-adaptive method which uses CM-LLR+MAP [3]. For the baseline method, we used the whole adaptation utterances instead of only the vowels. Although the baseline method adapts not only the spectral envelope but all of the acoustic features (log-f0, MVF and duration), we restricted the comparison to the MCEP model adaptation. As for the log-f0 model, we applied the average correction explained in subsection 3.4 both for the proposed method and for the baseline method. The MVF and duration models were simply not adapted.

4.2. Perceptual test and results

For the evaluation, 10 short sentences were synthesised for each adapted speaker and method, i.e. 120 sentences. A total of 25 evaluators took part in the evaluation, 6 of them being experts in speech technologies. Two sentences per speaker and method were randomly selected from the whole evaluation set and presented to each evaluator, so each evaluator rated 24 sentences. For every synthetic sentence, a recording of the corresponding original voice was also given as reference. The evaluators were asked to rate both the quality of the adapted voice (regardless of the quality of the original reference) and the similarity to the original target, both using a 1-5 scale. The mean opinion scores are shown in figures 1 and 2.

As can be seen from the figures, the results vary significantly depending on the speaker. These differences are due mainly to the varying quality of the recordings (note that the recordings were made at the user’s home). For M1, M2 and F2, the quality of the recordings is high and both adaptation methods achieve similar results in terms of quality. For M3, F1 and F3, the recordings contained audible environmental noise, and even some reverberation in the case of M3. This seems to be the reason of the low quality scores achieved by the baseline method for these three speakers, whereas the proposed method is robust against these phenomena thanks to the use of FW+AS transforms. This is a remarkable advantage when the recording environment is not controllable, as it is the case of our web interface [22]. Regarding similarity, the baseline method performs slightly better. This is logical, given that (i) the generic linear transforms involved in CMLLR are much more flexible than the FW+AS transforms, and (ii) the baseline method uses full utterances for adaptation. However, our method achieves relatively good similarity scores using only vowels for adaptation, which makes it suitable a priori for some types of pathological voices as we planned. From a different point of view, patients can be expected to prefer a roughly adapted voice with higher quality than the opposite, as long as basic characteristics of their voice (those that reflect their gender, age, etc.) are well captured.

5. Conclusions and future work

As a first step towards the goal of personalizing speech synthesizers using pathological voices, this paper has presented an adaptation method that requires only vocalic segments as input data. It is based on frequency warping plus amplitude scaling transforms, which can be applied to phonemes other than vowels without altering their phonetic content substantially. The results of the performed perceptual test show that the method provides high-quality adapted voices even when the recordings used as adaptation data have medium/low quality. Also, despite the use of vocalic segments only, its performance in terms of similarity to the target speaker is relatively good in comparison with state-of-the-art methods. Hence, the method has interesting properties for situations where the quality of the recordings is not guaranteed and the quality of the synthetic speech takes priority over a perfect capture of the target speaker’s identity. The next step is applying this method to specific types of dysarthric voices. Future works will also tackle the automatic detection of useful vocalic segments from pathological voice recordings.

6. Acknowledgements

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness (SpeechTech4All project, TEC2012-38939-C03-03) with FEDER support, and the Basque Government (Ber2tek project, IE12-333).
7. References


