Abstract

Voice Conversion (VC) is a task of transforming an utterance of a source speaker so that it is perceived as if spoken by a specified target speaker. A typical requirement in a VC system is to have a set of utterances recorded by both the speakers (called as parallel data), which is not always feasible. Further, in a cross-lingual voice conversion system, where the source and the target speaker’s language is different, it is impossible to have a parallel set of utterances. Hence, it is important to design an algorithm which performs a source speaker independent training. We propose a framework which captures speaker-specific characteristics and thus avoid the need for any training utterance from the source speaker. The proposed framework exploits the mapping abilities of Artificial Neural Networks (ANN) to estimate the conversion function. Experimental results reveal that the quality of the transformed speech is intelligible and has the characteristics of the target speaker.

1 Introduction

A Voice Conversion (VC) system transforms the utterance of a source speaker so that it is perceived as if spoken by a specified target speaker. Such a transformation involves mapping of spectral, excitation and prosodic features including duration, $F_0$ patterns of source speaker onto the target speaker’s acoustic space (Kain and Macon, 2001; Rao, 2009; Toth and Black, 2008). Several techniques have been proposed since the first code book based spectral transformation developed by (Abe et al., 1988). These techniques include use of mapping code books (Abe et al., 1988), Artificial Neural Networks (ANN) (Watanabe et al., 2002; Narendranath et al., 1995; Desai et al., 2009), dynamic frequency warping (Valbret et al., 1992) or Gaussian Mixture Model (GMM) (Kain and Macon, 2001; Stylianou et al., 1995; Toth and Black, 2007; Toda et al., 2004), etc. As the vocal tract shape between two speakers is non linear, ANN based method was proposed due to their ability to perform non-linear mapping (Narendranath et al., 1995). However, referring to the state-of-the-art systems, we find that GMMs are most widely used. Hence, a comparison between ANN and GMM based spectral transformation in the context of VC was performed and the results are provided in (Desai et al., 2009). They concluded that ANN based spectral mapping in voice conversion performs as good as the GMM based system.

Most VC methods rely on the existence of parallel data between the source and the target speakers (Abe et al., 1988; Kain and Macon, 2001; Narendranath et al., 1995; Stylianou et al., 1995; Toda et al., 2007; Toda et al., 2006), i.e, the source and the target speakers record a same set of utterances. Availability of such parallel data enables to arrive at a relationship between utterances of source and target speakers at a phone/frame level and build a VC system which learns a transformation from say phone /a/ of the source speaker to phone /a/ of the target speaker. In a cross-lingual voice conversion system, as the source and the target speaker language is different, there is no possibility of recording the same set of utterances. However, clustering techniques could be used to derive a relationship between utterances of the source and the target features at phone or sound level (Ye and Young, 2004; Sundermann et al., 2006; Sunder-
get speaker could be clustered into K clusters using VQ techniques, then these K clusters could be used to annotate the utterances of source speaker to derive the relationship between source and target speakers at the cluster level. Such data could be referred to as pseudo-parallel data which could be used to build a cross-lingual voice conversion system.

While these methods avoid the need for parallel data, they still require speech data (non-parallel data) from the source speakers apriori to build the conversion models. This is a limitation to an application where an arbitrary user intends to transform his/her speech to a pre-defined target speaker without recording anything apriori. Thus it is worthwhile to investigate conversion models which captures speaker-specific characteristics of a target speaker and avoid the need for speech data from source speaker for training. Such conversion models not only allow an arbitrary speaker to transform his/her voice to a pre-defined target speaker but also find applications in cross-lingual voice conversion.

2 ANN model for voice conversion

Artificial Neural Network (ANN) models consist of interconnected processing nodes, where each node represents the model of an artificial neuron, and the interconnection between two nodes has a weight associated with it. ANN models with different topologies perform different pattern recognition tasks. For example, a feed-forward neural network can be designed to perform the task of pattern mapping, whereas a feedback network could be designed for the task of pattern association. A multi-layer feed forward neural network is used in this work to obtain the mapping function between the source and the target vectors. Figure 1 shows the architecture of a four layer ANN used to capture the transformation function for mapping the source speaker onto the target speaker’s acoustic space.

If \( G(x_t) \) denotes the ANN mapping of \( x_t \), then the error of mapping is given by

\[
\epsilon = \sum_t \| y_t - G(x_t) \|^2
\]

(1)

where \( G(x_t) \) is defined as

\[
G(x_t) = \tilde{g}(w^{(3)} g(w^{(2)} g(w^{(1)} x_t))),
\]

(2)

and

\[
\tilde{g}(\vartheta) = \vartheta, g(\vartheta) = a \tanh(b \vartheta).
\]

(3)

Here \( w^{(1)}, w^{(2)}, w^{(3)} \) represents the weight matrices of first, second and third hidden layers of ANN respectively. The values of the constants \( a \) and \( b \) used in tanh function are 1.7159 and 2/3 respectively. A generalized back propagation learning (Narendranath et al., 1995) is used to adjust the weights of the neural network so as to minimize \( \epsilon \), i.e., the mean squared error between the desired and the actual output values. Selection of initial weights, architecture of ANN, learning rate, momentum and number of iterations are some of the optimization parameters in training an ANN (Yegnarayana, 2004). Once the training is complete, we get a weight matrix that represents the mapping function between the source and the target speaker spectral features. Such a weight matrix can be used to transform a source feature vector to a target feature vector.

![Figure 1: Figure showing an architecture of a four layered ANN with N input and output nodes and M nodes in the hidden layers.](image)

3 Models capturing speaker-specific characteristics

The idea in building an ANN model to capture speaker-specific characteristics is as follows. Let \( l_q \) and \( s_q \) be two different representations of the speech signal from a target speaker \( q \). A mapping function \( \Omega(l_q) \) could be built to transform \( l_q \) to \( s_q \). Such a function would be specific to the speaker and could be considered as capturing the essential speaker-specific characteristics. The choice of representation of \( l_q \) and \( s_q \) play an important role in building such mapping networks and in their interpretation. If we assume that \( l_q \) represents linguistic information, and \( s_q \) represents linguistic and speaker information, then a mapping function from \( l_q \) to \( s_q \) should capture speaker-specific information in the process. The interpretation of order
of Linear Prediction (LP) could be applied in deriving $l_q$ and $s_q$. A lower order ($\leq 6$) LP spectrum captures first few formants and mostly characterizes the message (or linguistic) part of the signal, while a higher order ($\geq 12$) LP spectrum captures more details in the spectrum and hence captures message and speaker characteristics (Misra et al., 2003). Thus $l_q$ being represented using lower order LP spectrum of first few formants could be interpreted as speaker independent representation of the speech signal, and $s_q$ represented using traditional MCEPs could be interpreted as carrying message and speaker information. An ANN model could be trained to minimize the error $||s'_q - s_q||$, where $s'_q = \Omega(l_q)$.

In this work, $l_q$ is represented by six formants, their bandwidths and delta features and $s_q$ is represented by traditional MCEPs as it would allow us to synthesize using MLSA synthesis code. An ANN model is trained to map $l_q$ to $s_q$ using backpropagation learning algorithm. Once the model is trained, it could be used to convert $l_r$ to $s'_q$ where $l_r$ could be from any arbitrary speaker $r$.

Figure 2 and 3 show the block diagram of the training and testing module of the proposed approach as discussed in this section.

![Block Diagram of Training Module](image1)

**Figure 2:** *Figure showing the block diagram of the training module in the proposed method.*

![Block Diagram of Testing Module](image2)

**Figure 3:** *Figure showing the block diagram of the testing module in the proposed method.*

To validate our proposed algorithm, we conduct three tests.

1. We use parallel data (a set of 40 utterances from a female-male pair) where the formant related features are extracted from the source female speaker data and the MCEPs are extracted from the target male speaker data. We then use ANN to train on this data which not only tunes the parameters and architecture of ANN but also justify if MCEPs can be predicted from formant related features.

2. We extract formant related features and MCEPs for the target speaker data and train an ANN on this data. ANN configuration tuned in step 1 is used and the models built are used to transform an utterance from an arbitrary source speaker of the same language so that the performance in case of intra-lingual voice conversion can be noted.

3. The ANN model obtained in step 2 is used to perform cross-lingual voice conversion, where the source speaker utterance given to transform is of a language different that the target speaker’s language.

### 3.1 Database

The work presented in this paper is carried out on CMU ARCTIC database consisting of 7 speakers. Each speaker has recorded a set of 1132 phonetically balanced utterances (Kominek and Black, 2004). The database includes utterances of SLT (US Female), CLB (US Female), BDL (US Male), RMS (US Male), JMK (Canadian Male), AWB (Scottish Male), KSP (Indian Male).

We initially test the validity of the proposed algorithm on a parallel data set as explained in section 3.3. It should be noted that for a voice conversion system based on parallel data, about 30-50 utterances are needed to build voice conversion model (Toth and Black, 2007). Thus, for each speaker we took around 40 utterances as training data and a separate set of 59 utterances as testing data. Once the ANN parameters are set, we test the validity of our proposed algorithm which requires only the target speakers data (explained in section 3.4).

The formants, bandwidths, $F_0$ and probability of voicing which represent $l_q$ are extracted using ESPS toolkit (ESPS, ). The formants also undergo a normalization technique such as vocal tract length normalization explained in Section 3.4.1. 25 Mel-cepstral coefficients (MCEPs) which represent $s_q$ are extracted for every 5 ms.

### 3.2 Evaluation

Given an input vector of formant related features, ANN predicts MCEPs. Hence, we need to evaluate how close the predicted MCEPs are to the original target speaker’s MCEPs. Mel Cepstral Distortion (MCD) is an objective error measure used which is known to have correlation with the subjective test results (Toda et al., 2004). Thus MCD
is used to measure the quality of voice transformation (Toth and Black, 2007). MCD is related to filter characteristics and hence is an important measure to check the performance of mapping obtained by ANN/GMM network.

\[
MCD = (10/ln10) \times \sqrt{2 \times \sum_{i=1}^{25} (mc_i^t - mc_i^e)^2}
\]

where \(mc_i^t\) and \(mc_i^e\) denote the target and the estimated MCEP, respectively.

### 3.3 Experiments with Parallel Data

As an initial experiment, we used parallel data of BDL and SLT. Features representing \(l_r\) were extracted from BDL speaker and were mapped onto \(s_q\) of SLT. This experimentation was done mainly to tune the mapping process with respect to representation of \(l_r\) and hence the results could be compared with that obtained in (Desai et al., 2009). Training was done to map BDL formants to SLT MCEPs with only 40 utterances. Testing was done on a set of 59 utterances. As the experiments in this section represent a mapping between two different speakers, VTLN was not used. Table 1 shows different representations of \(l_r\) and their effect on MCD values. These different representations include combination of different number of formants and their bandwidths, delta and acceleration coefficients of formants and bandwidths, pitch and probability of voicing. From the results provided in Table 1 we can observe that experiment 9 (which uses six formants, six bandwidths, probability of voicing, pitch along with their delta and acceleration coefficients) employing an error correction network provided better results in terms of MCD values.

### 3.4 Experiments Using Target Speaker’s Data

In this work, we built an ANN model which maps \(l_q\) features of SLT onto \(s_q\) features of SLT. Here \(l_q\) is represented by formants, bandwidth, \(F_0\), probability of voicing and their delta coefficients as shown in Table 1. \(s_q\) is represented by MCEPs. The formants also undergo Vocal Tract Length Normalization (VTLN) process to compensate for the speaker effects.

#### 3.4.1 Vocal Tract Length Normalization

VTLN is a speaker normalization technique that tries to compensate for the effect of speaker-dependent vocal tract lengths. VTLN compensates for the speaker-dependent vocal tract lengths by warping the frequency axis of the magnitude spectrum of speech frames before deriving the features. Apart from the use of VTLN in speech recognition, they have also been used in voice conversion (Sundermann et al., 2004; Sundermann et al., 2003; Sundermann et al., 2006).

The sizes of speakers’ vocal tracts is a physiological factor which is directly related to the vocal tract’s resonant frequencies or formants. VTLN is utilized to predict the vocal tract length of a speaker and appropriately rescale the frequency axis of the spectral representation. i.e. VTLN aims at warping the frequency axis of a speech frame to change formants and bandwidths that highly depend on vocal tract (Sundermann et al., 2006). Following the work in (Faria, 2003), we estimate the warp factors using pitch information and modify both formants and bandwidths. There are three warping factors estimated with pitch proposed in (Faria, 2003). We have performed a few experiments and found that for our case use of piece-wise linear warping function is better as provided in equation below.

\[
f_i' = \begin{cases} 
  kf_i & : f_i \leq F_0 \cdot f_N \\
  kF_0 \cdot f_N - \frac{kF_0 \cdot f_i - kF_0 \cdot f_N}{f_N - F_0 \cdot t} & : F_0 \cdot t < f_i < f_N 
\end{cases}
\]

where \(k = 1 - 0.002(F_0 \cdot i - f_{\text{mean}})\), \(f_i\) is the formant / bandwidth frequency of frame \(i\) to be normalized, \((F_0 \cdot i)\) is the pitch value of the frame \(i\) and \(f_N\) is the sampling frequency. \(f_{\text{mean}}\) is the mean pitch of the target speaker.

#### 3.4.2 Error correction network

We introduce a concept of error correction network which is essentially an additional ANN network, used to map the predicted MCEPs to the target MCEPs so that the final output obtained features represent the target speaker in a better way. The block diagram for error correction network is shown in Figure 4. Once \(s_q'\) are obtained, they are given as input to second ANN and it is trained to reduced the error \(\|s_q' - s_q\|\). Such a network acts as error correction network to correct any errors made by first ANN. Let \(s_q''\) denote the output from error correction network. It is observed that while
the MCD values of $s_q'$ and $s_q''$ do not vary much, the speech synthesized from $s_q''$ was found to be smoother than that of speech synthesized from $s_q'$. To train the error correction network, we use 2-D features i.e., feature vectors from 3 left frames, and 3 right frames are added as context to the current frame. Thus the ANN model is trained with 175 dimensional vector (25 dimension MCEPs * (3+1+3)). The architecture of this error correction network is 175L 525N 525N 175L.

![Figure 4: A block diagram of an error correction network](image)

3.4.3 Results

Initial experiments on the proposed approach are carried out on parallel set of utterances where $l_q$ represented by $F_0$, probability of voicing, formants, bandwidths and their delta coefficients as shown in feature set for experiment 9 in Table 1 are extracted from BDL utterances. $s_q$ is represented by MCEPs extracted from SLT utterances. The formants and bandwidths undergo VTLN to normalize speaker specific characteristics. We use the concept of error correction network to improve the smoothness of the converted voice.

Table 2 provides the results for mapping $l_q$ (where $r$ = SLT, BDL, RMS, CLB, JMK, AWB, KSP voices) onto the acoustic space of SLT. To perform this mapping the voice conversion model is built to map $l_q$ to $s_q$ (where $q$ = SLT) is used. To perform VTLN, we have used the mean pitch value of SLT. Hence all the formants of source speaker are normalized with VTLN using mean of SLT $F_0$ and then are given to ANN to predict the 25 dimensional MCEPs. Similar results where the conversion model is built by capturing BDL speaker-specific features are provided in Table 3.

3.5 Subjective Evaluation

Experiments conducted to capture speaker-specific characteristics predicts MCEPs from formant-related features. However, to synthesize a speech signal from MCEPs we also need excitation information. Excitation signal in form of fundamental frequency is used and also needs to be transformed. When an arbitrary source speakers utterance is to be transformed, formant-related features and fundamental frequency information are extracted. The formant-related features are used to predict MCEPs as described in Section 3.4 and fundamental frequency is transformed as explained below.

A logarithm Gaussian normalized transformation (Liu et al., 2007) is used to transform the source speaker $F_0$ to target speaker $F_0$ as indicated in the equation below.

$$\log(F_{0\,\text{conv}}) = \mu_{\text{tgt}} + \frac{\sigma_{\text{tgt}}}{\sigma_{\text{src}}} (\log(F_{0\,\text{src}}) - \mu_{\text{src}})$$

(6)

where $\mu_{\text{src}}$ and $\sigma_{\text{src}}$ are the mean and variance of the fundamental frequencies in logarithm for the source speaker, $F_{0\,\text{src}}$ is the pitch of source speaker and $F_{0\,\text{conv}}$ is the converted pitch frequency.

Predicted MCEPs along with transformed $F_0$ can be used as input to Mel Log Spectral Approximation (MLSA) (Imai, 1983). Listening tests are performed whose results are provided in Table 4.

<table>
<thead>
<tr>
<th>Transformation speaker pair</th>
<th>MCD [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLT to SLT</td>
<td>3.966</td>
</tr>
<tr>
<td>BDL to SLT</td>
<td>6.153</td>
</tr>
<tr>
<td>RMS to SLT</td>
<td>6.650</td>
</tr>
<tr>
<td>CLB to SLT</td>
<td>5.405</td>
</tr>
<tr>
<td>JMK to SLT</td>
<td>6.754</td>
</tr>
<tr>
<td>AWB to SLT</td>
<td>6.758</td>
</tr>
<tr>
<td>KSP to SLT</td>
<td>7.142</td>
</tr>
</tbody>
</table>

Table 2: Performance of voice conversion model built by capturing speaker-specific features of SLT, i.e., by mapping $l_q$ to $s_q$, where $q$ = SLT.

<table>
<thead>
<tr>
<th>Transformation speaker pair</th>
<th>MCD [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDL to BDL</td>
<td>4.263</td>
</tr>
<tr>
<td>SLT to BDL</td>
<td>6.887</td>
</tr>
<tr>
<td>RMS to BDL</td>
<td>6.565</td>
</tr>
<tr>
<td>CLB to BDL</td>
<td>6.444</td>
</tr>
<tr>
<td>JMK to BDL</td>
<td>7.023</td>
</tr>
<tr>
<td>AWB to BDL</td>
<td>7.017</td>
</tr>
<tr>
<td>KSP to BDL</td>
<td>7.444</td>
</tr>
</tbody>
</table>

Table 3: Performance of voice conversion model built by capturing speaker-specific features of BDL, i.e., by mapping $l_q$ to $s_q$, where $q$ = BDL.
with MOS scores ranging from 1 to 5 and 5 being provided to the best utterance. A similarity test is also performed where the transformed utterances and the target speaker’s utterances are provided to the listener. The question asked was how close are the two speakers and the results of the same are provided in Table 4. The range of similarity test is also from 1 to 5 where 5 indicates that both the recordings are from the same speaker. For the listening tests we chose 3 utterances randomly from each of the transformation pairs. Table 4 provides a combined output of all speakers transformed to target speaker (SLT / BDL). There were 14 listeners who participated in the evaluations tests. The MOS scores and similarity test results are averaged over 14 listeners. A MOS score of 5 means that the transformed speech is as natural as spoken language and in similarity test, a score of 5 implies that the transformed speaker and the target speaker are the same.

Table 4: Subjective evaluation of voice conversion models built by capturing speaker-specific characteristics

<table>
<thead>
<tr>
<th>Target Speaker</th>
<th>MOS</th>
<th>Similarity tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDL</td>
<td>2.926</td>
<td>2.715</td>
</tr>
<tr>
<td>SLT</td>
<td>2.731</td>
<td>2.47</td>
</tr>
</tbody>
</table>

The results shown in Tables 2, 3 and 4 indicate that voice conversion models could be built by capturing speaker-specific characteristics using ANN models. As this approach do not need any utterances from source speaker to train a voice conversion model we can use this type of model to perform cross-lingual voice conversion.

3.6 Application to cross-lingual voice conversion

Table 5: Subjective results of cross-lingual transformation done using conversion model built by capturing speaker-specific characteristics. 10 utterances from each of Telugu (NK) and Hindi (PRA) speakers are transformed into BDL male speaker’s voice

<table>
<thead>
<tr>
<th>Source Speaker</th>
<th>Target Speaker</th>
<th>MOS</th>
<th>Similarity test</th>
</tr>
</thead>
<tbody>
<tr>
<td>NK(Telugu)</td>
<td>BDL(English)</td>
<td>2.88</td>
<td>2.77</td>
</tr>
<tr>
<td>PRA(Hindi)</td>
<td>BDL(English)</td>
<td>2.62</td>
<td>2.15</td>
</tr>
</tbody>
</table>

Cross-lingual voice conversion is a task where the language of the source and the target speakers is different. In the case of speech-to-speech translation system where a source speaker does not know the target language but wants to convey information in his/her voice in the target language, cross-lingual voice conversion assumes important. As we now know that the availability of parallel data is very important to be able to learn the transformation, it is also important to note that this task is difficult in the case of cross-lingual voice conversion. One solution to such issues is to perform a unit selection approach (Sundermann et al., 2004; Sundermann et al., 2003; Sundermann et al., 2006) to find units in target speaker utterances that are close to the source speaker or use utterances recorded by a bi-lingual speaker (Jian and Yang, 2007; Mouchtaris et al., 2004). Our solution to cross-lingual voice conversion is to employ the ANN model which captures speaker-specific char-
acteristics.

In this context, we performed an experiment to transform two female speakers (NK, PRA) speaking Telugu and Hindi respectively into a male voice speaking English (US male - BDL). Our goal here is to transform NK, PRA voice into BDL voice and hence the output will be as if BDL speaking in Telugu and Hindi respectively. We make use of BDL models built in Section 3.4 to capture speaker-specific characteristics.10 utterances from NK, PRA voices were transformed into BDL voice and we performed MOS test and similarity test to evaluate the performance of this transformation. Table 5 provides the MOS and similarity test results averaged over all listeners. There were 10 native listeners of Telugu and Hindi who participated in the evaluations tests. The MOS scores in Table 5 indicate the transformed voice was intelligible. The similarity tests indicate the cross-lingual transformation could be achieved using ANN models, and the output is intelligible and possess the characteristics of BDL voice.

4 Conclusion

In this work, we have exploited the mapping abilities of ANN to perform voice conversion. Having justified that there is a need for an algorithm which trains on only the target speaker data, we were able to show that it is possible to build such a model. The proposed approach uses ANN to capture the speaker-specific characteristics and such an approach does not require any speech data from source speaker. We were also able to show that the proposed approach could be applied to cross-lingual conversion with no extra efforts. The output of the proposed approach is intelligible and has the characteristics of the target speaker.

Acknowledgment

The authors would like to thank E. Veera Raghavendra, Keri Venkatesh and Sachin Joshi of Speech Lab, IIIT-Hyderabad for useful discussions and suggestions. We also thank all the people who participated in various listening tests carried out for this work.

References


