1. ABSTRACT

Voice transformation, for example, from a male speaker to a female speaker, is achieved here using a two-level dynamic warping. An outer warping process, which temporally aligns blocks of speech (dynamic time warp), invokes an inner warping process, which spectrally aligns based on magnitude spectra (dynamic frequency warp). The mapping function produced by the dynamic frequency warp is used to move spectral information from a source speaker to a target speaker. Information obtained by this process is used to train an artificial neural network to produce spectral warping output information based on spectral input data.

2. INTRODUCTION

Voice transformation refers to the process of changing the parameters of the speech or changing voice personality, to convert the speech uttered by one speaker (source speaker) to sound as if other speaker (target speaker) had spoken it[1]. Voice transformation has applications such as text-to-speech synthesis (TTS), international dubbing, health-care, multi-media, language education, music, and preprocessing for speech recognition, etc. [2, 3, 4]. In conventional voice transformation acoustic parameters (such as pitch and formant information) of the speech signal are collected from the source and target speaker voices to find a transformation map that be used to map the parameters of the source speaker onto the target speaker signal[5]. Altering formant frequencies by moving the pole locations of the related formant produces changes in the vocal tract model. That movement played as an important factor in voice transformation[6]. Extraction, modified and mapping pitch also considered as an important factor in voice transformation[7]. The approach taken here avoids the need to find acoustic parameters (e.g., pitch or formant model), and instead deals directly with spectral information. The transformation is accomplished using a two-level dynamic warp (DW). Based on the two-Level DW it is straightforward to map the source speech to target speech when both are available. But if the target speech is already available saying the desired target sentence, why bother with the transformation? A more challenging, but realistic, setting is when the target is not available saying the desired statement. Thus, a second phase of the research is to train an artificial neural network to produce the spectral warping function from only the source speaker information, based upon which the source speech may be warped to the target speech.

3. DYNAMIC WARPING

Dynamic time warping (DTW) is a method can be used to find an optimal alignment between the sequences of feature vectors from two different sources[8]. DTW has been used for a wide range of applications such as speech recognition [9], human motion animation [10], human activity recognition [11] and processor cache analysis [12]. We have extended DTW to a two-level mode, where the outer level seeks temporal alignment, using a norm computed using an inner frequency warping to align spectra. This has been used as a tool to help train caregivers working with individuals with speech disabilities [8].

It is obvious from the name “dynamic time warping” that DTW temporally aligns block of features to compensate for different speech rates. For example, temporal alignment between male speech segment and female speech segment by comparing certain features occurring at as a function of time. Dynamic frequency warping (DFW) or “inner warping” can also be used to perform spectral alignment based on spectral magnitudes of blocks of speech data, such as aligning spectral features of male speaker to spectral features of female speaker. In this paper we applied the combination of inner and outer warping, simply referring as “dynamic warping,” or DW.

Our explanation for the DW starts with a brief summary of the outer warping (DTW). For two speakers \(i (i = 1, 2)\), let \(s_i(\ell, m)\) denote the feature at the temporal block number \(\ell\) and frequency index \(m\). The vector \(s_i(\ell, :)\) is a (spectral) feature vector at block \(\ell\) of user \(i\). A speech signal for speaker \(i\) (such as a spoken sentence) is represented as a sequence of \(T\), feature vectors computed from overlapping blocks of speech information

\[
S_i = \{s_i(1,:) , s_i(2,:) , \ldots , s_i(T_i, :)\}, \quad i = 1, 2,
\]

Let \(d(s_1(\ell_1 , :), s_2(\ell_2 , :))\) denote a metric distance between vectors \(s_1(\ell_1 , :)\) and \(s_2(\ell_2 , :)\) at a specific block times \(\ell_1\) and \(\ell_2\). This metric is computed using DFW, as explained below. DTW is employed to find the distance between the sequence \(S_1\) and the sequence \(S_2\). The notation \(d_T(\ell_1, \ell_2)\) represents the minimal cost between the sequences \(S_1\) and \(S_2\), up to block numbers \(\ell_1\) and \(\ell_2\), allowing some alignment motion between blocks of the two different users. (The subscript \(T\) refers to time warping.) \(d_T(\ell_1, \ell_2)\) is recursively computed as

\[
d_T(\ell_1, \ell_2) = d(s_1(\ell_1 , :), s_2(\ell_2 , :)) + \\
+ \min\{d_T(\ell_1 - 1, \ell_2), d_T(\ell_1, \ell_2 - 1), \}
\]

At the end of that specific portion of speech \((\ell_1 = T_1, \ell_2 = T_2)\), time warping produces an overall warped metric distance
$d_T(T_1, T_2)$ between the sequences $S_1$ and $S_2$, which may be represented as $d_T(S_1, S_2)$. This is suggested in Figure 1. This DTW accounts for temporal shifts due to the rate of speaking differences between the two speakers. The output of this DTW is the temporal alignment of the source signal with respect to the target signal.

As mentioned above, the distance $d(S_1(\ell_1,:), s_2(\ell_2,:))$ can be obtained through warping process between these feature vectors. This warping is called dynamic frequency warping (DFW) or inner warping. Frequency warping can be calculated in a way similar to the time warping. Let $s_i(\cdot) = s_i(\ell_i, :)$, $i = 1, 2$ denote the spectral information at timeslice $\ell_i$ that is passed in from the outer warping (DTW) to the inner warping (DFW). DFW is applied to find the distance between $s_1(\cdot)$ and $s_2(\cdot)$ as

$$d_S(k_1, k_2) = \text{dist}(s_1(k_1), s_2(k_2)) +$$

$$+ \min[d_S(k_1 - 1, k_2), d_S(k_1, k_2 - 1),$$

$$d_S(k_1 - 1, k_2 - 1)].$$

Here the subscript $S$ refers to a frequency warping, and $\text{dist}(s_1(k_1), s_2(k_2))$ is represent the metric distance between elements of the spectral vectors. In our case, we use spectral feature vectors computed as the positive frequency components of FFT of windowed data. The metric distance used looks only at spectral magnitude, so

$$\text{dist}(s_1(k_1), s_2(k_2)) = ||s_1(k_1)|| - ||s_2(k_2)||.$$ At the end of frequency warping process, the distance between spectral vectors is computed as

$$d(s_1(\ell_1,:), s_2(\ell_2,:)) \triangleq d_S(K, K),$$

where $K$ is the number of elements in each spectral feature vector.

DFW also produces a sequence of indices $a = (a(1), a(2), \ldots, a(M))$ and $b = (b(1), b(2), \ldots, b(M))$, which are called the warping function paths, such that the new spectral vectors $s_1(\ell_1, a(\cdot))$ and $s_2(\ell_2, b(\cdot))$ are as similar as possible (e.g., peaks and valleys of $s_1$ align with peaks and valleys, respectively, of $s_2$). The source speaker $s_1$ spectrum is transformed to match the target speaker $s_2$ by creating a modified source speaker $s_1$ according to

$$\hat{s}_1(b(i)) = s_1(a(i)), \quad i = 1, 2, \ldots, M.$$ This spectral alignment map drags spectral components of source blocks to produce transformed speech in the frequency domain. This data is inverse Fourier transformed and added in sequence to produce the transformed signal. After warping, filtering is performed to mitigate signal processing artifacts.

The length M of the warping function paths may vary from one spectral feature vector to another, depending on how many turns the spectral warping paths have.

The combination of the two warping processes is illustrated in Figure 2. Starting from the speech signal at the bottom of the diagram, the speech signal is split into different overlapping segments. Feature vectors (spectral information) for each segment was computed using an FFT. These spectral features information are passed through DTW for temporal alignment, where at every stage of the temporal alignment, spectral alignment or warping is employed using DFW.

4. SPECTRAL FEATURES

Experiments were carried out on CMU ARCTIC database, which consist of seven speakers (two US females, two US males, one Canadian male, one Scottish male and one Indian male). Each speaker has recorded a set of 1132 phrases. In experiments presented here one US male and one US female are chosen to do the voice transformation from male to female.

Speech data was sampled at 16000 samples/sec. These feature vectors that were used in the DW were considered to be the positive-frequency spectral information in the frequency domain, calculated using the FFT. Each sentence was temporally segmented into 32-ms segments, using a Hamming window with 16-ms overlapping, zero-padded, then transformed using a 512-point FFT. The $K = 256$ positive frequency spectral elements in the frequency domain were used as a feature vector.

5. SPECTRAL WARping USING ANN

The two-level dynamic warping procedure described above was used to obtain training data to train an artificial neural network (NN). The NN input data was spectral feature vectors from the source speaker. The NN output data was spectral warping.
path information a and b, which can be used to do the spectral warping function.

Training data is obtained as portrayed in Figure 4, as follows: Source and target information are aligned using DTW. The warping paths a and b computed by the DW when it calls the inner spectral warping are saved as training data. Backpropagation learning is applied to adjust the weights matrix of the desired network such as to minimize the mean square error between the calculated and actual spectral warping paths (a and b). In this work, different NN structures with different parameters were examined.

The length of the warping paths a and b may be different from one segment of speech to another. In order to have the neural network be able to have a constant output length (without zero-padding the output data, which would introduce warping artifacts) the a and b information are interpolated to produce a modified a and b which is the same (maximum) length for all segments of the speech vector. This warping is computed as described here using an example.

- Let $a = (1, 2, 3, 4, 5)$ (length of a is 6), which has the indices of $[1, 2, 3, 4, 5, 6]$. (That is, 1-based indexing.) Let the new length is $\max_{index} = 10$. ($\max_{index}$ is the maximum length of vector a in the training data set.)
- Interpret the indices as 0-based, so the new indices are $[0, 1, 2, 3, 4, 5]$.
- Create a new set of interpolated indices (generally with non-integer values) as
  \[ \tilde{n} = \lfloor 0 : (\max_{index} - 1) \rfloor \times \frac{\text{length}(a) - 1}{\max_{index} - 1} \]
  In the example, the new indices are $\tilde{n} = [0, 0.556, 1.111, 1.667, 2.222, 2.778, 3.333, 3.889, 4.444, 5]$.
- The new indices still fall within the range of 0 to 5, but there are now $\max_{index}$ index values.
- Interpolate the a values using the interpolated indices as
  \[ \tilde{a} = \text{interpolate}(a, [0 : (\text{length}(a) - 1)], \tilde{n}) \]
  In this example the new $\tilde{a}$ values are $\tilde{a} = [1, 1.583, 2.103, 2.638, 3.204, 3.857, 4.066, 4.003, 4.1506, 5]$.

The above process can be repeated for function path b to find $\tilde{b}$. Figure 5 illustrates normal function paths for a and b with interpolated paths. ($\tilde{a}$ and $\tilde{b}$ respectively.)

In order to get the original length of the warping paths a and b, reverse interpolation was applied to the estimated $\tilde{a}$ and $\tilde{b}$ obtained by testing the trained neural network. This reversing was applied as described here using an example.

- Let $\tilde{a} = (1, 1.583, 2.103, 2.638, 3.264, 3.857, 4.066, 4.003, 4.1506, 5)$ (length of $\tilde{a}$ is 10). Suppose the original length of a is $\text{length}(a) = 6$ (as in the previous example).
- Create a new set of reverse interpolated indices as
  \[ \hat{n}_{rev} = \lfloor 0 : (\text{length}(a) - 1) \rfloor \times \frac{\text{length}(\tilde{a}) - 1}{\text{length}(a) - 1} \]
  In the example, the new reversing indices are $\hat{n}_{rev} = [0, 1.8, 3.6, 5.4, 7.2, 9]$.
- Reverse interpolation of the $\hat{a}$ values using reverse interpolated indices are
  \[ \hat{a}_{rev} = \text{interpolate}(\hat{a}, [0 : (\text{length}(\tilde{a}) - 1)], \hat{n}_{rev}) \]
  In this example the new $\hat{a}_{rev}$ values are $\hat{a}_{rev} = (1, 2.0003, 3.0038, 3.9916, 3.9993, 5)$.

The above process can be repeated for function path $\tilde{b}$ to find $\hat{b}_{rev}$. Once these interpolated values are computed, then the warped spectrum is computed by created an interpolated spectrum according to

\[ s_1(\hat{b}_{rev}(i)) = s_1(\hat{a}_{rev}(i)), \quad i = 1, 2, \ldots, M. \]
After identifying the starting point and ending point of the first 500 phrases of CMU ARCTIC database (with recordings of US male and US female speakers), DW was applied on the spectral magnitudes feature vectors that are extracted from the FFT for each 32-ms segment of speech with 50% overlap windowed using a Hamming window. Temporal alignment (DTW) was done on the source spectral feature vectors with respect to the target male signal. (The temporal aligned source spectral feature vectors were saved for later use within the NN.) DFW was applied to do the spectral alignment and to find the sequence of path indices (a and b). These indices were used to transform the source speaker $s_1$ spectrum to match the target speaker $s_2$.

Different NN structures were applied (such us number of layers, number of neurons) to do a voice transformation. After trying several architectures, we chose to use two hidden layers with 500 and 1000 neurons for the first and second layers, respectively. Neural network computations were performed using TensorFlow in Python.

7. RESULTS

Figure 3 shows a typical spectrogram for voice transformation from male speaker to female speaker using the phrase “Author of the danger trail, Philip Steels, etc.” from the ARCTIC database using the training data (known male and female speaker information). Figure 6 shows the spectral feature information for one segment of speech for female with the time aligned spectral warped segment for male. By the DFW, the peaks and valleys of the male spectrum are aligned to the locations of the peaks and valleys of the female spectrum. Figure 7(a) shows a spectrogram warped information of Neural Net result as a voice transformation from male speaker to female speaker. Acoustically, the transformed signal is clearly looks like the previous result. Figure 7(b) shows the mean square error. Figure 8 shows how the NN has learned a warping path (the a warping data) at various segments of the speech signal after 4000 training iterations. The DW path is shown in blue and the NN-learned path is in orange.

8. CONCLUSION

The DW warping method introduced here produces female-sounding speech from male speech, but with some signal processing artifacts. This is also the case when the warping information is provided by a trained NN.

Future work will focus on how to improve the mapping by exploring and eliminating the causes of the signal processing artifacts, and exploring different NN architectures. Another item of interest is whether this warping method can be used to produce different language accents.

9. REFERENCES

[1] E. Helander, J. Schwarz, J. Nurminen, H. Silen, and M. Gabouj, “On the impact of alignment on voice conversion performance,” International Speech Communication Association, vol. 1, pp. 1453–1456, Sept. 2008.

[2] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, “Transformation of formants for voice conversion using artificial neural networks,” Speech Communication, vol. 16, pp. 207–216, Nov. 1995.

[3] O. Türk, O. Büyük, A. Haznedaroglu, and L. M. Arslan, “Application of voice conversion for cross-language rap singing transformation,” International Conference on Acoustics, Speech and Signal Processing, pp. 3597–3600, May 2009.

[4] Y. Stylianou, “Voice transformation: a survey,” 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3585–3588, May 2009.

[5] X. Weichao and Z. Linghua, “Vocal tract spectrum transformation based on clustering in voice conversion system,” International Conference on Information and Automation, pp. 236–240, June 2012.

[6] H. Kuwabara and T. Takagi, “Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method,” Speech Communication, 1991.

[7] J. Slifka and T. R. Anderson, “Speaker modification with lpc pole analysis,” International Conference on Acoustics, Speech, and Signal Processing, 1995.

[8] A.-W. Al-Dulaimi, S. Budge, S. A. Borrie, T. K. Moon, and J. H. Gunther, “A tool for training speech imitation accuracy,” in 2018 52nd Asilomar Conference on Signals, Systems, and Computers, pp. 1086–1090, IEEE, 2018.

[9] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, pp. 43–49, Feb. 1978.

[10] E. Hsu, K. Pulli, and J. Popovic, “Style translation for human motion,” ACM Transactions on Graphics, vol. 24, pp. 1082–1089, July 2005.

[11] K. Kulkarni, G. Evangelidis, J. Cech, and R. Horaud, “Continuous action recognition based on sequence alignment,” International Journal of Computer Vision, pp. 1–32, June 2014.

[12] J. Zhao and L. Itti, “ShapeDTW: Shape dynamic time warping,” Pattern Recognition, vol. 74, pp. 171–184, Feb. 2018.
Fig. 6. One Spectral Segment Feature for Male, Female and Warped Male to Female

Fig. 7. Neural Net Results

Fig. 8. Warping Path Comparison