Accent Conversion

Yidong Fu
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA, 15213
yidongf@andrew.cmu.edu

Linbin Luo
Electrical and Computer Engineering
Carnegie Mellon University
Mountain View, CA, 94035
linbinl@andrew.cmu.edu

WaiChau Sze
Electrical and Computer Engineering
Carnegie Mellon University
Mountain View, CA, 94035
waichaus@andrew.cmu.edu

Zixuan Wang
Electrical and Computer Engineering
Carnegie Mellon University
Mountain View, CA, 94035
zixuanwa@andrew.cmu.edu

1 Introduction

Accent Conversion (AC), a subset of Voice Conversion (VC), is proved to be very important and widely used in language learning. Researches have shown that the best starting point for a beginner to learn a second language is to imitate the pronunciation patterns (or accent) of a native speaker [1],[2]. In this process, an AC algorithm can generate speech with different accents from the speakers' voice. It can significantly boost second language learning.

Furthermore, although accent is a critical factor in speech recognition, most speech recognition systems are accent-biased since they are trained on "standard" accents. A study showed that even the most sophisticated speech recognizer has a higher word error rate of around 16% when the speech is in the same context but with different accents [3]. Since Accent Conversion has been proved to have the ability to transform non-native speech to sound as if the speaker had a native accent, this research introduces it as a light but efficient tool to narrow this gap.

To solve the aforementioned speech recognition weakness when encountering speech with different accents, this project aims to study and compare different methods to robustly convert English speeches with various accents in a non-deep-learning manner.

2 Related Work

A foreign accent can be defined as deviations from the expected acoustic (e.g. formants) and prosodic (e.g. intonation, duration, and andrate) norms of a language. In contrast with VC, AC seeks to combine the linguistic content and pronunciation characteristics of the source speaker with the voice quality of the target speaker.

Early attempts on AC used voice morphing [4]-[7] to control the degree of accent by blending spectral components from the native and non-native speakers. In [8, 9], the authors used PSOLA to modify the duration and pitch patterns of accented speech.

A couple of studies also tried to blend native and non-native spectra to control the accent. Huckvale and Yanagisawa[10] used an English text-to-speech (TTS) system to simulate English-accented Japanese utterances. Aryal and Gutierrez-Osuna [11] adapted VC techniques, replacing Dynamic Time Warping (DTW) with a technique that matched source and target frames. Later, Zhao [12] used PPG similarity instead of MFCC similarity to pair acoustic frames.

In this study, we will research different AC methods and verify their ability to boost speech recognition performance on different accents. Reconstructions and experiments with existing accent conversion

methods would be a perfect starting point for this project. Since most researches have been done on VC, we will also research if we can transfer some of them to solve this problem. After we explore methods like sparing coding and GMM, advantages and disadvantages of each method would be analyzed, and we expect to improve these methods to maximize the performance of speech recognition on different accents.

3 Dataset

• CMU_ARCTIC speech synthesis databases: The databases consist of around 1150 utterances selected from out-of-copyright texts from Project Gutenberg. Speakers are native English speakers, and other accented speakers such as Canadian, Scottish and Indian. We currently used two speakers from the CMU_ARCTIC corpus: BDL (English male), and KSP (Indian male). For each speaker, we now selected two utterances from each speaker for experiment. The sources from this dataset are .wav speech files (16kHz), and the targets are also the .wav speech files after accent conversion. Pre-processing of our data is introduced in our Working Experiment section.

4 Approach

As aforementioned, Accent Conversion is a sub-task of Voice Conversion, thus, the team adopted methods from Voice Conversion, combined with Time Alignment technique, and utilized some signal processing approaches to construct speech without accent. The figure below 4 demonstrates an overall workflow of the project.

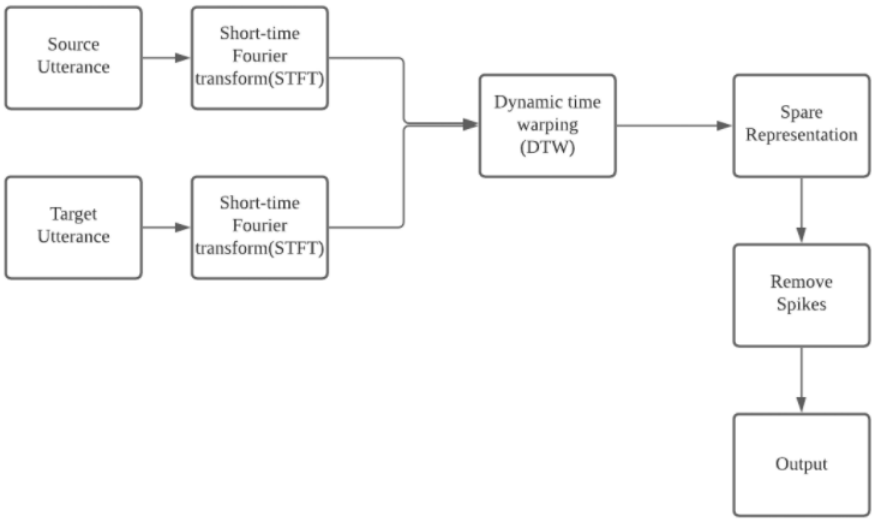

Figure 1: Accent Conversion Method Workflow

Short Time Fourier Transform (STFT) is adopted as the main feature extractor. It is a Fourier-based transform which can give the sinusoidal frequency as well as the phase content of a small window of a signal. It is a widely used method in the domain of Voice Conversion and it is proved to be instrumental in Accent Conversion. It provides a time-localized frequency information in the cases that a signal's frequency components changes all the time. Different from traditional Fourier Transform which determines the frequency information that are averaged over the complete time interval of a signal, STFT maintains more time-based frequency information. The team also explored the Mel-Frequency Cepstrum (MFC) comparing to STFT and the ablation study can be found in the experiment and result section. MFC is a widely used representation especially in sound processing. MFCCs are a set of coefficients that construct an MFC. It is based on the linear cosine transform of a log power spectrum on a nonlinear met scale of frequency.

On top of the feature extractor, it is crucial to apply time alignment technique to correctly match the source frames and target frames. The team used Dynamic Time Warping to realize time alignment. The DTW algorithms can align two time-dependent signals that are similar but locally out of phase. It aims to find an optimal warping alignment path between the source and the target by minimizing the overall cost of the cost matrix. A cost matrix is a matrix with each entry representing the Euclidean distance between each pair of the elements of the source sequence and target sequence.

Sparse Representation of extracted features are then applied. The figure below 4 demonstrates a flow chart of dictionary learning. In the training part, Source data and Target data are concatenated, then a general Dictionary is studied based on the combined data. The learning of the n components dictionary is to optimize the following equation:

$$\begin{aligned} &argmin_{D,W}(\frac{1}{2}||Y - DW||_{Fro^2} + \alpha||W||), \\ &with\ ||D_k||_2 = 1, for\ k = 0, 1, \ldots, n-1, \end{aligned} \tag{1}$$

where Fro stands for the Frobenius norm. The Dictionary would later be split to generate Dictionary D_s and D_t (Source Dictionary and Target Dictionary). In testing stage, a weight W is found by sparse coding source Data matrix by D_s. This method is completed based on LASSO:

$$argmin_W(\frac{1}{2}||Y - DW|| + \alpha||W||). \tag{2}$$

During the optimization, α will help control the sparsity of the result. At last, by using the same W, the matrix multiplication of W and D_t will be the converted target data matrix.

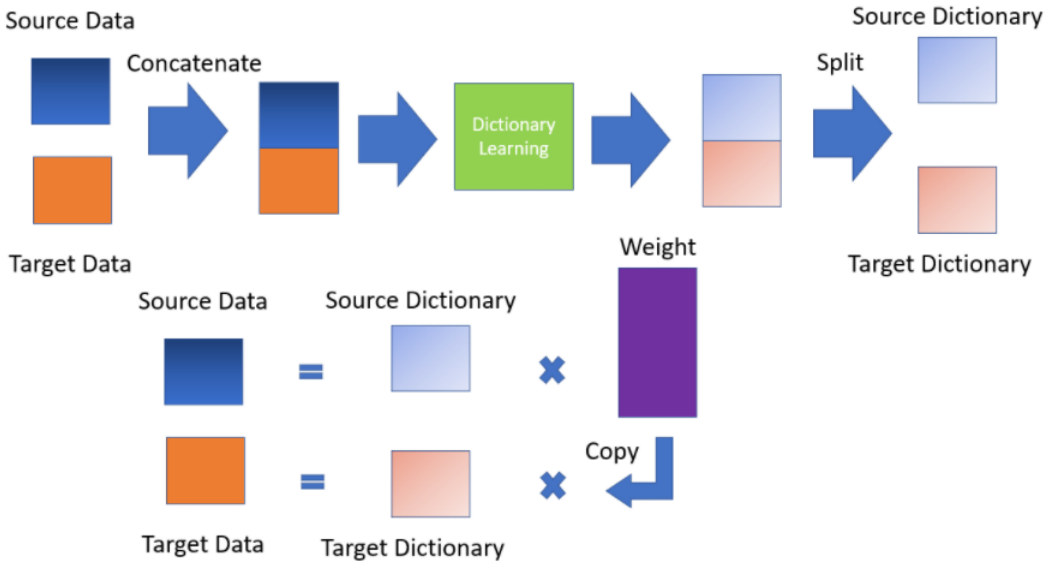

Figure 2: Dictionary Learning Workflow

Ultimately, we removes the spikes of the constructed accent speech with mean sliding window and median sliding window. The mean sliding window replaces the sequence frame with the average value of a fixed size window of the signal. The median sliding window, as its name suggested, replaces the sequence frame with the median value of a fixed size window of the signal.

5 Evaluation Method

Current evaluation metrics named Mel-Cepstral Distortion is used to examine the synthesis quality. A lower MCD means that the source conversion is closer to the ground-truth target.

$$MCD = \frac{10\sqrt{2}}{\ln 10}\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\sqrt{\sum_{i}(\mathrm{C_{ti}} - \hat{\mathrm{C}}_{\mathrm{ti}})^2}, \tag{3}$$

where t represents the time frame index and i represents the MFCCs index.

6 Experiment and Result

6.1 MFCCs + DTW

We extracted the Mel Frequency Cepstral Coefficients (MFCCs) features of source and target utterances, and then applied time-aligned algorithm using DTW based on the MFCCs features. We then reconstructed the time-aligned results to audio file. We finally compared the reconstruction audio with the source audio using Mel-Cepstral Distortion evaluation metric method.

6.2 STFT + DTW

Similar to 6.1 experiment, we replaced the MFCCs features with Short Time Fourier Transformation features. That is to say, we applied the DTW algorithm based on STFT features, and then completed the reconstruction work, also with evaluation.

6.3 STFT + DTW + Sparse Coding

Based on the results of the 6.2, we continued to apply Sparse Coding (SC) algorithm using DictionaryLearning api from scikit-learn package. Because the MFCCs + DTW experiment result is worse than STFT + DTW, we did not execute the MFCCs features with SC.

6.4 STFT + DTW + Sparse Coding + Remove Spikes

After 6.3 experiment, we found that the reconstruction audio contains some spikes, so we tried to apply an algorithm having been described in Section 4 Approach to remove the spikes to improve reconstruction results.

6.5 Results

Refer to the introduction of our experiments, we carried out four experiments with different features and further steps on CMU ARCTIC Dataset and finally conduced ablation study on their Mel-Cepstral Distortion. The results can be found in the table 1 below.

	Original	MFCCs + DTW	STFT + DTW	STFT + DTW + SC	STFT + DTW + SC + Remove Spikes
MCD	659.17	603	554	432.3	314.9

Table 1: Mel-Cepstral Distortion of Different Methods.

The ablation study proves that the methods would add up to decent results. It can be clearly seen from the table that Sparse Coding leads to a significant drop in the Mel-Cepstral Distortion. Feature extraction method such as MFCCs and STFT have relatively closer result but STFT appears to be better. The MFCCs method does not seem to provide performance as we expected. The generated voice sounds a little broken and a guess for this would be that the performance of time alignment method is not consistent. Another interesting finding is that the final output after removing spikes does not sound as good as that before removing spikes. However, the MCD does show that the discrepancy between source and target is closer.

7 Discussion and Analysis

In this project, Short Time Fourier Transform (STFT) is adopted as the main feature extractor, which can give the sinusoidal frequency as well as the phase content of a small window of a signal. Sparse Representation itself performed not ideally in Accent Conversion because it only focuses on pairing voice characteristics(i.e. pitch range, pronunciation, and speaking rate), but linguistic content is also

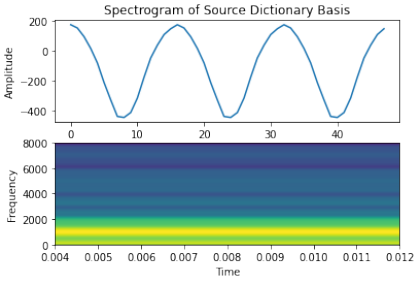

Figure 3: Source Speech

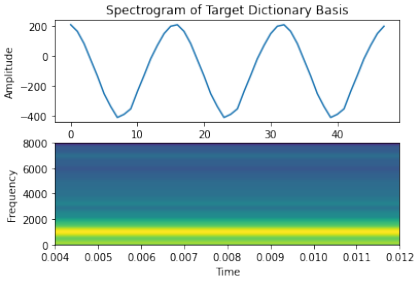

Figure 4: Target Speech

important. Time Alignment solves the linguistic problem like pairing the length of vowel on temporal aspect partially, but requires target pair. Removing spike makes output more smooth, and improves result quality.

We explored the similarity of the basis from the sparse representation of spectrogram and the phases. The sparse representation of spectrogram and the phases share the same weight. Figures 3 and 4 demonstrates the spectrogram of one paired basis of a certain frame from the source domain and the target domain and they are very similar.

In this model, we made use of DTW time alignment to improve the model performance. However, this means that for each source audio, we needed a target audio in order to complete the time alignment. In the future, we considered eliminating the time alignment part from testing part of the models.

References

[1] Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari, “Voice conversion using sequence-to-sequence learning of context poste- rior probabilities,” Interspeech, pp. 1268–1272, 2017.

[2] Das, A., Zhao, G., Levis, J., Chukharev-Hudilainen, E., & Gutierrez-Osuna, R. (2020). Understanding the Effect of Voice Quality and Accent on Talker Similarity.

[3] Koenecke, Allison, et al. "Racial disparities in automated speech recognition." Proceedings of the National Academy of Sciences 117.14 (2020): 7684-7689.

[4] Felps, D., Bortfeld, H. and Gutierrez-Osuna, R., 2009. Foreign accent conversion in computer assisted pronunciation training. Speech communication, 51(10), pp.920-932.

[5] Aryal, S., Felps, D. and Gutierrez-Osuna, R., 2013. Foreign accent conversion through voice morphing. In Interspeech (pp. 3077-3081).

[6] D. Felps and R. Gutierrez-Osuna, "Developing objective measures of foreign-accent conversion," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 5, pp. 1030-1040, 2010.

[7] M. Huckvale and K. Yanagisawa, "Spoken language conversion with accent morphing," in ISCA Speech Synthesis Workshop, 2007, pp. 64-70.

[8] S. Zhao, S. N. Koh, S. I. Yann, and K. K. Luke, "Feedback utterances for computer-adied language learning using accent reduction and voice conversion method," in ICASSP, 2013, pp. 8208-8212.

[9] J. Jügler, F. Zimmerer, J. Trouvain, and B. Möbius, "The perceptual effect of L1 prosody transplantation on L2 Speech: The case of French accented German," in Interspeech, 2016, pp. 67-71.

[10] M. Huckvale and K. Yanagisawa, “Spoken language conversion with accent morphing,” in Sixth ISCA Workshop on Speech Synthesis, 2007, pp. 64–70.

[11] S. Aryal and R. Gutierrez-Osuna, “Can voice conversion be used to reduce non-native accents?” in Proc. ICASSP, May 2014, pp. 7879– 7883.

[12] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent Conversion Using Phonetic Posteriorgrams," in ICASSP, 2018, pp. 5314-5318.

[13] S. Ding, G. Zhao, R. Gutierrez-Osuna, “Learning Structured Sparse Representations for Voice Conversion.” IEEE Xplore, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8910392.

[14] D. J. Berndt and J. Clifford, “Using dynamic time warping to find patterns in time series,” in Proc. Int. Conf. Knowl. Discovery Data Mining, 1994, pp. 359–370.

[15] S. Liu et al., "End-To-End Accent Conversion Without Using Native Utterances," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6289-6293, doi: 10.1109/ICASSP40776.2020.9053797.