Voice Conversion using Joint Density Non-Negative Matrix Factorization

Sean Pogorelc
Carnegie Mellon University
spogorel@andrew.cmu.edu

Arjun Raguram
Carnegie Mellon University
arjunraguram@cmu.edu

Samuel Rainey
Carnegie Mellon University
sir@cmu.edu

Zacchaeus Williams
Carnegie Mellon University
zvw@andrew.cmu.edu

1 Team Member Names

Sean Pogorelc, Arjun Raguram, Samuel Rainey, Zacchaeus Williams

2 Introduction

This project implements a voice conversion (VC) system to convert a speech utterance from a source speaker into the corresponding speech utterance of a target speaker. Applications of this work include recovering speech from a source speaker with vocal damage and emulating a voice that is in demand such as a voice actor/actress. This project uses Joint Dictionary NMF (JD-NMF) to implement VC.

3 Related Work

We base our work on a VC paper by Fu et al [1]. They propose an augmented JD-NMF algorithm to reduce computational complexity as compared to traditional NMF approaches. Their proposed method suggests using a jointly-trained dictionary to reduce the number of bases that need to be trained using NMF, a core component of VC. This approach requires combining the source and target spectograms when finding bases, requiring them to have the same dimensionality. See Figure 1 below:

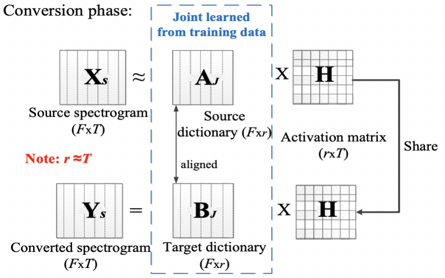

Figure 1: JD-NMF Diagram

Fu et al. made other substantial improvements by incorporating Dynamic Time Warping (DTW) to align the frames of source and target voices. We previously looked into utilizing the method discussed in Aihara et al. [2], which utilized a sparse dictionary of phoneme bases, selected based on a Gaussian Mixture Model (GMM) algorithm. We decided not to incorporate a GMM algorithm due to its need for a large training data set, which nullifies the benefit gained from JD-NMF.

4 Datasets

Two team members, Samuel Rainey (Sam) and Zacchaeus Williams (Zacchaeus), recorded 20 sentences each for the dataset. The JD-NMF approach for voice conversion pairs well with this method of data preparation since it requires less training data than traditional methods, consequently reducing the number of recordings to be made. Twenty sentences of varying length and phonetic structure were chosen to create a diverse set of data that would be more likely to enable instances of successful VC. Sentences were chosen from CMU_ARCTIC [4], select poems, and a collection of phonetic pangrams (sentences with all 44 phonemes from English included). Both Sam and Zacchaeus recorded audio in the same bedroom environment at roughly the same time of day, same distance from the microphone, and same amplification gain, using a Shure SM7B microphone fed into a Zoom H6 audio interface connected to a laptop running Audacity via USB, using a sampling rate of 44.1kHz and a bit depth of 16 bits. Sentences were recorded with similar, but often differing inflections and prosody to reduce variance in bases. In addition, sentences were recorded with timing alignment in mind. As such, Zaccheaus recorded all of the sentences, and then Sam recorded each sentence while listening to Zaccheaus' recording. Achieving approximate alignment was a major consideration in recording our own data set. Each sentence recording was trimmed to match length between speakers and normalized to a peak of -0.1dB using Audacity. A link to the dataset, including a Word document of training utterances, can be found here (click link).

5 Results

We used a combination of both quantitative and qualitative metrics to judge the performance of our system, favoring qualitative metrics for potential practical use considerations. The quantitative metrics include SDR (Source to Distortion Ratio) and SAR (Source to Artifact Ratio). SDR determines reconstruction error whereas SAR measures the presence of artifacts, which have no relation to the source itself. Since our conversion is a combination of two different sound files, the SDR and SAR is computed for comparison between the conversion and both the source and target files.

NFFT	SDR (Source)	SAR (Source)	SDR (Target)	SAR (Target)
256	-13.334	-13.334	-27.082	-27.082
512	-25.748	-25.748	-32.303	-32.303
1024	-31.089	-31.089	-34.339	-34.339
2048	-34.646	-34.646	-34.934	-34.934

Equations for reference:

$$SDR = \log \frac{||S_{target}||^2}{||e_{artif} + e_{interf} + e_{noise}||^2}$$

$$SAR = \log \frac{||S_{target} + e_{interf} + e_{noise}||^2}{||e_{artif}||^2}$$

Observe the effect of the number of windowed samples used in the Fast Fourier Transform algorithm (NFFT) in computing the Short Time Fourier Transform (STFT) on SDR and SAR. As NFFT increases from $256 \rightarrow 2048$, SDR and SAR decrease. Note that there is no difference between the computed SDR and SAR, indicating that all of the distortion in the voice conversion is a result of unwanted artifacts introduced from the JD-NMF algorithm. However, SDR and SAR do not completely characterize the efficacy of voice conversion. The SDR and SAR values for all NFFT sizes are low, whereas a good audio synthesis should produce a positive SDR and SAR result, but both are negative when looking at the conversion compared to both the source and target audio. Looking just at these metrics, the conclusion is to use a small NFFT size, 256, but with negative SDR and SAR values, a small NFFT size for voice conversion can lead to poor results for qualitative evaluation.

Since we are dealing with audio being used for practical purposes, such as increasing intelligibility of patients who have endured an oral surgery, we opted to favor qualitative metrics for the evaluation of the resulting voice conversion over the previously examined quantitative ones. The assessment centered around three metrics: speech pattern match, intelligibility, and naturalness. Speech pattern match evaluates whether the utterance follows the inflection, pacing, and prosody of the source speaker while using the natural voice qualities of the target speaker. Intelligibility evaluates whether the words and meaning of the utterance is clear and interpreted as intended by the source speaker. Naturalness evaluates whether the VC sounds like the target speaker originally speaking the utterance rather than being an algorithmically generated utterance.

To start, in contrast to the suggestions of the quantitative analysis, qualitatively we found that the best results were produced using an NFFT size of 1024. As we decreased the NFFT size from 1024 we found the audio, while louder, had a quality to it that sounded like it was being spoken underwater. This effect largely disappears for the NFFT size of 1024 and then by 2048 we can start to here an echo forming as the target's voice slightly separates from the source's speech pattern. For the rest of the analysis a NFFT size of 1024 was used.

In terms of the conversions following the source's speech pattern, we found that our conversion was successful as when we listened to both the original target and source audio, we found the conversion's pacing and rhythm to be a closer match to the source than the target. We could especially hear this in the pacing of audio, as when we played either the target or source followed by the conversion, it was easy to distinguish that the source's pacing matched the conversion's.

In terms of the intelligibility of the conversions we had both group members and people not associated with the project listen to the audio and try to make out the phrase being said. For recording 14 (z14.wav and s14.wav; used for audio/visuals in report and presentation) we found that on average most people were able to make out the majority of words after only a handful of times hearing it, with the one missed most being the name 'Tudor' near the middle, due to unfamiliarity with the name. Those with prior knowledge of the phrase, or those told what the phrase was beforehand, had no issue identifying the words and comprehending the conversion. Of note is that this was our best conversion, likely due to it being decently well aligned and on the shorter side, which we acknowledge as showing that while our JD-NMF worked well, its performance relied heavily on the severity of misalignment of the source and target files.

Lastly is the metric of naturalness. This is the weakest metric of the 3 and the one where we can not say we feel like we succeeded. We would deem the conversions to sound 'natural' if it sounded like we had just loaded in a professional recording and played it, similar to how the source/target files sound. However, our conversion files have a very noticeable echo to them which makes it hard for us to call it a success in this category.

With all of these things considered we deemed that our JD-NMF algorithm was successful in converting source speech patterns to target speech, however, there is much room for improvement in terms of the recording quality and proper temporal alignment that could further improve these results beyond their current state. This is especially true in the case of severe temporal misalignment, as our lack of proper temporal alignment causes severely misaligned files to produce very poor quality conversions that fail all 3 of the aforementioned qualitative metrics.

6 Discussion and Analysis

Continuing with why our conversions have an echo, we deemed that this is likely attributable to the lack of proper alignment techniques pre-NMF (such as alignment using DTW). To compensate for this we increased the length of the windowed signal for each FFT in the STFT to give us worse temporal resolution, hoping to use larger temporal bins to mitigate the timing offsets between the files. We did this by manipulating NFFT since in the documentation for librosa.stft() they describe how the window length parameter is tied to NFFT, and is either set to the provided NFFT size or zero-padded to match it if a smaller window length is specified (trying to make it larger will result in an error). While this worked well enough to give us successful conversions, the system would likely work for a larger variety of utterances and might sound more natural if proper alignment is performed.

For a visual reference to the imperfect alignment, below is the "aligned" spectrogram used for the first NMF of recording 14 (referenced as our best recording in the results section) where we train the joint exemplar dictionary and shared activation matrix:

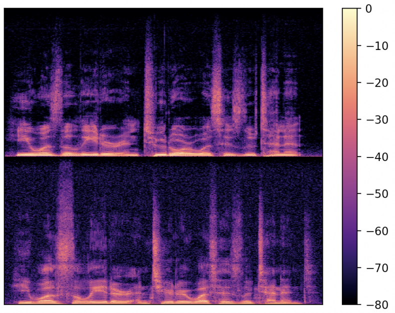

Figure 2: Spectogram for Recording 14: Target and Source (1024 NFFT)

For comparison here are the estimations for the source and target audio respectively using the calculated joint exemplar dictionary. Notice how they show the same structures as the above target and source spectrogram but are noticeably smeared across time due to the imperfect alignment:

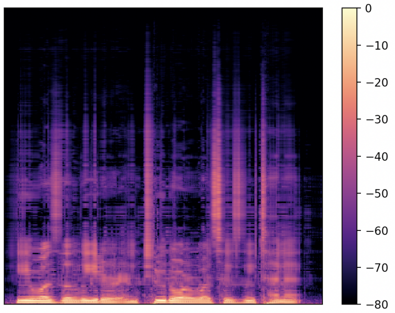

Figure 3: Spectogram for Recording 14: Estimated Source (1024 NFFT)

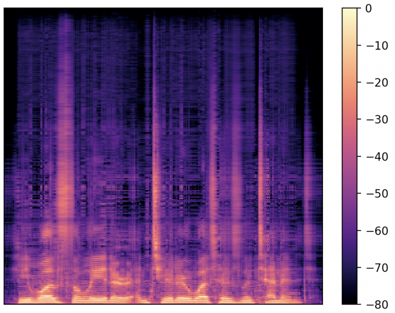

Figure 4: Spectogram for Recording 14: Estimated Target (1024 NFFT)

Now for comparison here is the recording's conversion spectrogram. Notice how it shows the same structure as the above estimated target and source spectrograms, including the smear across time due to the imperfect alignment:

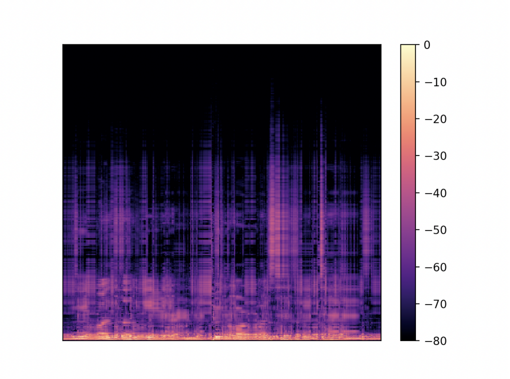

Figure 5: Spectogram for Recording 14: Result from VC (1024 NFFT)

In the case of this conversion our JD-NMF was able to give a good conversion of the source and target audio, but we attributed some of this success to the lack of severe misalignment which allowed our system to skip the typical alignment step. With proper alignment we feel like our result would be better, especially in the sense that the aforementioned 'echo' should be fixed as there would be less smearing of the converted audio in time. We also believe the produced audio would be of higher quality as we could use a smaller NFFT size, allowing for better temporal resolution in the STFT and therefore more precise measurement of the utterance in time.

The lack of temporal alignment was not purely negligence on our part however. We had tried to implement alignment using the warp path from dynamic time warping (DTW). The alignment algorithm that was considered inserts silence from a specified silent audio file to even out the length between the target and source audio files. Further, a crossfade algorithm was applied to smooth transitions between audio portions and inserted silence portions. While this algorithm is effective in aligning according to the warp path determined by DTW, there are two problems that we encountered, both relating to audio of a long duration (greater than 2 seconds). For such audio, the alignment algorithm needs to make a lot of adjustments, causing the audio to sound buzzy or choppy, even with the crossfade. Portions that need to be slowed down significantly sound downsampled. The larger problem is a flawed warp path determined by DTW. Comparing the specified warp path to the aligned files, alignment occurs where it is specified to happen by the warp path but the warp path does not always properly capture corresponding segments of audio. This follows as DTW is traditionally used for segments of audio that are similar but scaled in length, and it can fall short for speakers with different vocal features over a long duration. A solution that we must experiment with further is using shorter utterances, including just a word or two.

An additional portion of the JD-NMF that is worth mentioning is the number of iterations of NMF performed in each of the 2 separate instances. We found that while the first NMF that learns the joint exemplar dictionary and shared activation matrix required significantly more iterations to converge (~1000 iterations), the second NMF that retrains only the activation matrix based on the found source exemplar dictionary runs for far fewer iterations (~20). While we think this gap in iterations may be due in large part to the higher complexity of solving for the joint exemplar dictionary, our current number of iterations used for each NMF is based on a combination of trial and error and parameter discussions in the other papers read (mainly Fu et al.), thus leaving a potentially large amount of room for improvement in run time efficiency and/or conversion quality with additional experimentation.

Lastly we will examine why we believe the quantitative results proved to be misleading in their indication of success and which NFFT size was best. Given that our conversion is a combination of the source and target audio signals its no surprise that the resulting SDR and SAR were very low. However, when we examine the SDR and SAR of the bases we get the same trend of lower NFFTs having higher SDR/SARs, and the lower valued NFFTs have fairly good SDR/SAR values when compared to before.

NFFT	**SDR (Source)**	**SAR (Source)**	**SDR (Target)**	**SAR (Target)**
256	31.017	31.017	26.012	26.012
512	26.742	26.742	21.195	21.195
1024	-2.822	-2.822	-4.284	-4.284
2048	-8.286	-8.286	-14.252	-14.252

This indicates that quantitatively we get worse representations of the bases when using higher NFFT values which then results in worse approximations of the original source/target audio. This propagates through the use of the source bases in the 2nd NMF and the target's bases when finding the converted spectrogram for our converted audio. As to the cause of this we can likely attribute it to the NFFT trick used to solve the alignment issues, which is causing the estimated audio to be a slightly blurred version of the original (as seen in figures 3 and 4). Now as for why the results then are qualitatively better for the size 1024 NFFT than the smaller sizes, see the above discussion above detailing the effects of misalignment and how we worked to counter them with larger temporal windows.

7 References

[1] Fu, S.-W., Li, P.-C., Lai, Y.-H., Yang, C.-C., Li-Chen, amp; Tsao, Y. (2016). "Joint Dictionary Learning-based Non-Negative Matrix Factorization for Voice Conversion to Improve Speech Intelligibility After Oral Surgery." https://www.citi.sinica.edu.tw/papers/yu.tsao/5692-F.pdf.

[2] R. Aihara, T. Nakashika, T. Takiguchi and Y. Ariki, "Voice conversion based on Non-negative matrix factorization using phoneme-categorized dictionary," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 7894-7898, doi: 10.1109/ICASSP.2014.6855137.

[3] Lorenzo-Trueba, Jaime; Yamagishi, Junichi; Toda, Tomoki; Saito, Daisuke; Villavicencio, Fernando; Kinnunen, Tomi; Ling, Zhenhua. (2018). The Voice Conversion Challenge 2018: database and results, [sound]. The Centre for Speech Technology Research, The University of Edinburgh, UK. https://doi.org/10.7488/ds/2337.

[4] Black, Alan W. (2021). CMU_ARCTIC speech synthesis databases. Language Technologies Institute at Carnegie Mellon University. http://www.festvox.org/cmu_arctic/