Incorporating durational modification in voice transformation

Arthur Toth, Alan W Black

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
{atoth,awb}@cs.cmu.edu

Abstract

Voice transformation is the process of using a small amount of speech data from a target speaker to build a transformation model that can be used to generate arbitrary speech that sounds like the target speaker. One common current technique is building Gausian Mixture Models to map spectral aspects from source to target speakers. This paper proposes the use of duration models to improve the transformation models and output speech quality. Testing across seven target speakers shows a statistically significant improvement in a popular objective metric when duration modification is performed both during training and testing of a Gaussian Mixture Model mapping based voice transformation system.

Index Terms: voice transformation, speech synthesis

1. Introduction

Voice Transformation (VT) is the process of building a statistical model from a small amount of target speaker's speech and using it to convert existing (or new) speech from a different speaker such that it sounds like the target speaker. Voice transformation has a number of uses in fooling speaker ID systems, disguising voices, and entertainment.

There are a number of properties of a voice that distinguish it from others including: spectral, excitation, and prosodic aspects as well as higher level aspects such as word and subject choice.

Our particular interest is being able to quickly generate a speech synthesizer in the target speaker's voice. This would allow us to more quickly offer a large number of different voices, as well as be able to offer a particular voice without a large effort from the target speaker in recording data. Current corpus based synthesis techniques may require many hours of speech, and people without professional training find difficulty in delivering it consistently, though newer statistical parametric speech synthesis techniques [1] can be successful with much less data.

The system we use here first generates synthetic speech with an existing synthesizer. This removes the requirement to have parallel recordings for the source and target speaker, because the synthesizer can generate whatever the target speaker said. However typically the target speaker records a small set of appropriately balanced sentences.

For our source speaker in these tests we use a standard diphone voice [2], not because it is a high quality synthesizer, but because it is a very consistent synthesizer. Importantly it can easily control pitch and duration of synthesized speech using modification using a residual excited LPC technique [3].

One of the problems with typical voice transformation systems based on Gaussian Mixture Model (GMM) mapping is that they make naive assumptions about prosody. One particular issue is that the frame-by-frame mapping from source speaker to target speaker causes the transformed speech to have the same duration as the source speaker's speech. This is incorrect as the transformed speech should really have duration statistics that match those of the target speaker, and the target speaker will have different duration statistics from the source speaker. This paper investigates attempts to use duration statistics to modify both the training and testing processes for a GMM mapping based voice transformation system. Though the topic of modifying duration in the voice transformation testing phase has been investigated before (e.g. [4]), our approach differs in that we modify the source speaker durations as well, thus affecting the entire spectral conversion process.

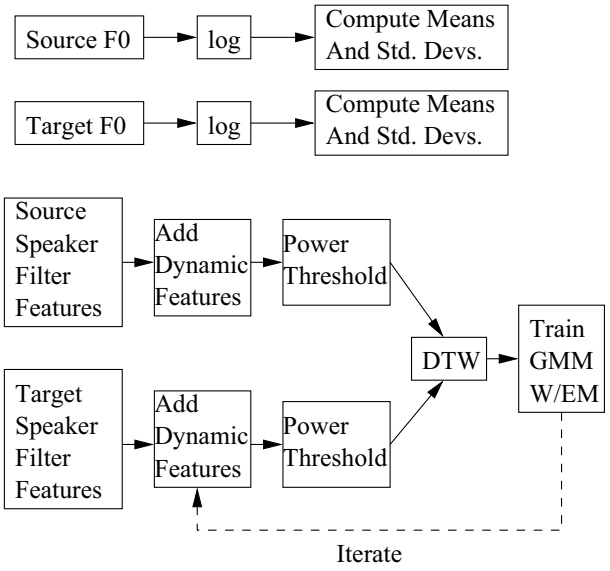

Figure 1: *Voice Transformation Training Process*

2. Baseline Voice Transformation System

Numerous methods for VT have been attempted over the past 20 years. Early techniques used vector quantization [5]. The one used in this paper is based on techniques created by Tomoki Toda [6], and is freely available in FestVox project [7] scripts. The Gaussian Mixture Model (GMM) mapping technique used in these scripts, which is based on earlier techniques [8] [9], is still one of the prevalent voice transformation techniques. The basic idea is that the distribution of acoustic features from the source and target speaker can be modeled with a GMM.

2.1. Training

The specific training process used in the FestVox voice transformation scripts is depicted in Figure 1 [10]. During training,

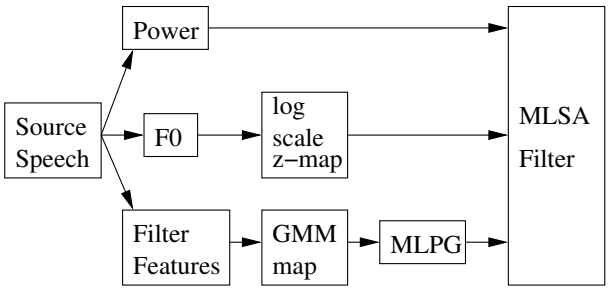

Figure 2: *Voice Transformation Process*

speech from the source and target speakers based on the same text is analyzed every 5ms, and F_0 estimates and filter features are created. The F_0 estimates are either positive values, which are fundamental frequency estimates, or zeros for speech that is judged to be unvoiced. The filter features are a representation of the spectral envelope called MCEPs, which approximate Mel-scale cepstra. Different processes are used to train maps between F_0 estimates and maps between filter features. For the F_0 maps, the means and standard deviations of the logarithms of the F_0 estimates on voiced speech are stored for the source and target speakers. The training of the map between the filter features is more complicated. The MCEPs are augmented with dynamic features which are derived by applying a short-term weighted window to the MCEPs. After power thresholding, Dynamic Time Warping (DTW) is used to align source speaker frames with target speaker frames based on these augmented feature vectors, and a GMM is used to model the joint probabilities. The DTW and GMM training steps are repeated 2 more times.

2.2. Testing

The transformation process is depicted in Figure 2 [10]. During transformation, a source speaker utterance is analyzed, and F_0 estimates and MCEPs are produced every 5ms. The source speaker F_0 estimates are transformed to target speaker F_0 estimates by taking their logarithms, performing a z-score mapping, and then exponentiating them. A z-score mapping from s to t is defined as

$$t = (s - \overline{s})\frac{\sigma_t}{\sigma_s} + \overline{t}$$

where $\overline{s}$ and $\overline{t}$ are the means of s and t, respectively, and σ_s and σ_t are the standard deviations of s and t, respectively. The source speaker filter features are transformed to target speaker filter feature estimates by augmenting them with the same type of dynamic features that were produced during training, and using maximum likelihood estimation based on the previously learned GMM to predict target speaker MCEPs. The resulting predictions for the F_0 and MCEP values are then used as inputs to a Mel Log Spectral Approximation (MLSA) filter [11] to produce synthetic speech.

3. Data

The data used for the target speakers in the following experiments was taken from the CMU ARCTIC databases [12] for the currently available 7 speakers. Table 1 lists some of the speakers' characteristics. For each speaker, the first 50 utterances from the A subset were used for training voice transformation models, and the 101st through 110th utterances from the A subset were used for testing voice transformation models.

ID	Gender	Dialect
awb	male	Scottish
bdl	male	American
clb	female	American
jmk	male	Canadian
ksp	male	Indian
rms	male	American
slt	female	American

Table 1: *CMU ARCTIC Speaker Characteristics*

For the source speaker utterances, the text for the same ARCTIC database utterances was synthesized by the kal-diphone synthesizer from the Festival Speech Synthesis System distribution [2].

4. Duration Modification Experiments

Different speakers speak at different rates and also vary in the relative lengths of the phonetic segments of their speech. Voice transformation should take such durational differences into account in order to better map from one speaker to another. Unfortunately, typical GMM mapping based voice transformation techniques ignore these differences because transformed speech is produced on a frame-by-frame basis from source speaker speech and thus has the duration characteristics of the source speaker, instead of the desired duration characteristics of the target speaker.

The following experiments attempt to address this issue by modifying duration characteristics of the source speaker utterances to make them match those of the target speaker during voice transformation training and testing. During training, it would also be possible to pursue the opposite strategy of modifying the target speaker utterance duration characteristics to match those of the source speaker, but this approach would require more effort. As we use a synthetic voice for a source speaker and are specifying duration targets anyway, it is simpler to modify the source speaker durations. During testing, the target speaker test utterances are not available, so it would not be possible to use them.

4.1. Training Duration Modification

In the following experiments, training duration modification is performed by the following process:

1. Synthesize the source speaker training utterances based on the default duration characteristics of the synthetic voice.

2. Use DTW between the source speaker training utterances and the target speaker training utterances to label the segment endpoints for the target speaker training utterances.

3. Resynthesize the source speaker training utterances using the same segment endpoints that were estimated from the DTW.

The resynthesized source speaker utterances from this process are then used with the target speaker utterances in the VT training process described in Section 2.1.

4.2. Testing Duration Modification

During testing, the transformed utterances produced by the baseline voice transformation system will have the same durations as the source speaker test utterances, due to the frame-by-frame conversion process. Our strategy to produce transformed utterances whose duration characteristics more closely match those of the target speaker is indirect. We modify the durations in the source speaker test utterances to be more like those from target speaker utterances, and the modification carries over to the transformed utterances due to the frame-by-frame conversion process.

During testing, it is not possible to perform the same duration modification procedure that was performed during training because the target speaker test utterances are not available, and DTW cannot be performed to estimate segment endpoints. Instead we tried a different procedure for modifying durations on the source speaker test utterances:

1. Collect average speech durations from the source and target speaker training utterances. These durations were based on utterances, not segments. Leading and trailing silences were not included in the durations.
2. Synthesize the source speaker test utterances based on the default duration characteristics of the synthetic voice.
3. Calculate new source speaker segment endpoints by multiplying the current ones by the average target speaker speech duration and dividing by the average source speaker speech duration.
4. Resynthesize the source speaker test utterances based on the new endpoints.

5. Evaluation

After creating new versions of the voice transformation process involving duration modification, it is necessary to compare them to the baseline and each other to determine whether there has been an improvement. This leads to questions of what is important in voice transformation, and how can it be measured.

Three qualities that are typically considered important in transformed speech are naturalness, intelligibility, and identity. Speech is natural if it sounds like it came from a person, it is intelligible if the words can be understood, and transformed speech has the proper identity if it sounds like the target speaker spoke it.

Popular methods of evaluating voice transformation can be divided into two categories: subjective tests and objective tests. Subjective tests involve having humans listen to examples of speech and rate them. The advantage of subjective tests is that they measure human perception directly, and human opinion is typically the standard. Some disadvantages of subjective tests are that they can be difficult to design, costly to implement, and some method must be devised to analyze the differing subjective opinions from different people. Objective tests involve calculating metrics automatically from the data without human intervention. Some advantages of objective tests are that they can be quick to perform and are automatic. The disadvantage of objective tests is that none of them appear to correspond perfectly with human judgment of transformed speech, which is typically the standard. However, some objective metrics, such as mel-cepstral distortion (MCD) appear to have a reasonable degree of correlation with human judgment of voice transformation quality [6].

Target	Baseline	Train Mod.	Test Mod.	Both Mod.
awb	6.12 (2.30)	6.15 (2.34)	5.87 (2.04)	5.85 (2.08)
bdl	7.62 (3.10)	7.56 (3.01)	7.66 (3.02)	7.55 (2.90)
clb	6.69 (2.33)	6.69 (2.34)	6.76 (2.41)	6.74 (2.42)
jmk	6.92 (2.50)	6.83 (2.43)	6.90 (2.46)	6.81 (2.40)
ksp	7.14 (2.28)	7.28 (2.35)	7.16 (2.31)	7.30 (2.40)
rms	6.75 (2.34)	6.79 (2.28)	6.75 (2.24)	6.76 (2.19)
slt	6.89 (2.54)	6.77 (2.42)	6.90 (2.52)	6.71 (2.25)
Avg.	6.88	6.87	6.86	6.82

Table 2: *VT Duration Modification Results: MCD means (MCD std. dev.)*

MCD is essentially a weighted Euclidean distance based on mel-cepstral feature vectors. It is defined by the formula:

$$MCD = \frac{10\sqrt{2}}{\ln 10}\sqrt{\sum_{d=1}^{24}(t_d - r_d)^2}$$

where t_d is the dth mel-cepstral coefficient of a frame of speech from a test speaker and r_d is the dth mel-cepstral coefficient of a frame of speech from a reference speaker. The smaller the MCD between two frames, the more similar they are. For utterances, average MCDs are calculated over frames.

There is an additional complication when using MCD to evaluate voice transformation. The durations of the transformed test utterances and the target speaker reference utterances are typically different, so some process has to be used to align the utterances before comparison. It is typical to use DTW for this purpose.

6. Results

Table 2 shows the results of the duration modification experiments. The "Target" column lists the tags for the ARCTIC database speakers who were the voice transformation targets. The "Baseline" column lists the MCD means and standard deviations for the voice transformation system described in Section 2. The "Train Mod." column lists the MCD means and standard deviations for voice transformation using the training duration modification procedure described in Section 4.1 with the baseline testing procedure. The "Test Mod." column lists the MCD means and standard deviations for voice transformation using the baseline training procedure, but the testing procedure described in Section 4.2. Finally, the "Both Mod." column lists the MCD means and standard deviations for voice transformation using both the training duration modification procedure and the testing duration modification procedure. The best average results came from performing duration modifications both during training and during testing.

Although the duration modification technique used during training that was described in Section 4.1 cannot be performed during testing due to the lack of information about target speaker test utterances, the simpler duration modification procedure used during testing that was described in Section 4.2 could be used during training. Using this procedure during training is a switch to a more naive duration model, because it only takes total speech length into account and does not incorporate further information about individual segment durations. A comparison of using these two different duration modification strategies during training while using the same testing duration modification process in both sets of trials is given in Table 3. The column

Target	Rate Mod.	Segment Mod.
awb	6.07 (2.02)	5.85 (2.08)
bdl	7.59 (2.98)	7.55 (2.90)
clb	6.84 (2.48)	6.74 (2.42)
jmk	6.81 (2.43)	6.81 (2.40)
ksp	7.20 (2.31)	7.30 (2.40)
rms	6.74 (2.26)	6.76 (2.19)
slt	6.83 (2.50)	6.71 (2.25)
Avg.	6.87	6.82

Table 3: *Comparison of Different Training Duration Modifications While Using the Same Testing Duration Modification: MCD means (MCD std. dev.)*

labeled "Rate Mod." uses the strategy from Section 4.2 during both training and testing. The column labeled "Segment Mod." uses the strategy from Section 4.1 during training and the strategy from Section 4.2 during testing. It is the same as the column marked "Both Mod." in Table 2. On average, using the more sophisticated duration modification strategy during training performed better than using the more naive one when both were used in combination with the same testing duration modification strategy.

7. Discussion

In Table 2, the best overall results came from using duration modification procedures during both training and testing, with a reduction in MCD of 0.06 from the baseline. The combination outperformed using duration modification only during training and only during testing. The difference between the baseline voice transformation and the use of duration modification during both training and testing was statistically significant, because a one-tailed, paired T-test based on the mean MCDs for each utterance gave a value of $p = 0.0043$. Duration modification appears to be a worthwhile addition to a typical GMM mapping based voice transformation process.

A number of directions for future work appear promising. One potential area of improvement is to incorporate segment statistics into the duration modification process used during testing. Currently, the duration modification during testing amounts to a speaker rate mapping, without any variation based on phonetic segment types. This is a bit of a mismatch with the duration modification during training, which is based on segment lengths. In order to use segmentation during testing, some problems need to be handled. One is that training sets for voice transformation are typically small and don't provide full phonetic coverage. A testing process that uses phonetic segment information would need a strategy for processing segments that didn't appear in the training set. One possibility is to cluster phones, perhaps based on acoustic-phonetic features, and share statistics within clusters. Another possibility is to back off to speaker rate mapping of durations when test segments are not present in the training data.

Another direction is to explore additional evaluation metrics. It is possible that some of the advantage of duration modification techniques is not captured by the average MCD metric due to the DTW smoothing out differences between durations. A comparison with subjective listening tests would be good.

Looking more broadly at the issues involved with GMM mapping based voice transformation shows that duration is only part of the prosody that could have a better mapping. Power and fundamental frequency should also be investigated. Power is currently handled by simply using the same 0th order mel-cepstral coefficient for the transformed speech that was extracted from the source speaker speech. The baseline voice transformation procedure makes no attempt to make it more similar to the power of the target speaker. Fundamental frequency is transformed by z-score mapping the source speaker estimates to the target speaker estimates. Although this approach can give reasonable global statistics for the fundamental frequency of the transformed speech, such as mean and standard deviation, it can miss subtler local differences between the source and target speakers involving trajectories. In general, there is significant room for improvement in the handling of prosody in voice transformation.

8. Acknowledgments

This work was supported by the US National Science Foundation under grant number 00414675 TRANSFORM: flexible voice synthesis through articulatory voice transformation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

9. References

[1] A. Black, H. Zen, and K. Tokuda, "Statistical parametric synthesis," in *ICASSP2007*, Hawaii, 2007.

[2] K. Lenzo and A. Black, "Diphone collection and synthesis," in *ICSLP2000*, Beijing, China., 2000.

[3] M. Hunt, D. Zwierynski, and R. Carr, "Issues in high quality LPC analysis and synthesis," in *Eurospeech89*, vol. 2, Paris, France, 1989, pp. 348–351.

[4] D. Rentzos, S. Vaseghi, Q. Yan, and C.-H. Ho, "Voice conversion through transformation of spectral and intonation features," in *ICASSP2004*, Montreal, 2004.

[5] M. Abe, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *J. Acoust. Soc. Jpn. (E)*, vol. 11, pp. 71–76, 1990.

[6] T. Toda, A. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis," in *5th ISCA Speech Synthesis Workshop*, June 2004.

[7] A. Black and K. Lenzo, "Building voices in the Festival speech synthesis system," 2000, http://festvox.org/bsv/.

[8] Y. Stylianou, O. Cappé, and E. Moulines, "Statistical methods for voice quality transformation," in *Proc. EUROSPEECH95*, Madrid, Spain, 1995, pp. 447–450.

[9] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *ICASSP-98*, vol. 1, Seattle, Washington, 1998, pp. 285–288.

[10] A. Toth and A. Black, "Using articulatory position data in voice transformation," in *ISCA SSW6*, 2007.

[11] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proceedings of ICASSP 83*, 1983, pp. 93–96.

[12] J. Kominek and B. A., "The CMU ARCTIC speech databases for speech synthesis research," Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-LTI-03-177 http://festvox.org/cmu_arctic/, 2003.