Using Articulatory Position Data in Voice Transformation

Arthur R. Toth, Alan W Black

Language Technologies Institute, Carnegie Mellon University,
Pittsburgh, PA, USA

atoth@cs.cmu.edu, awb@cs.cmu.edu

Abstract

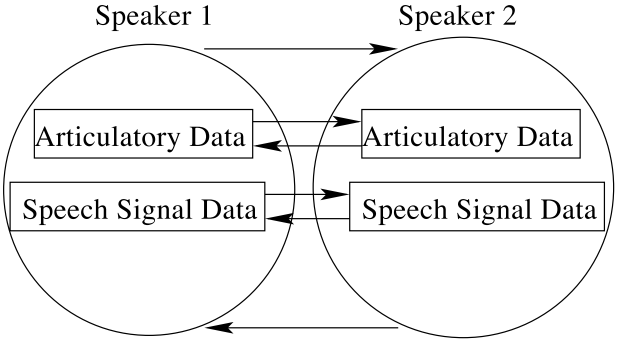
Articulatory position data is information about the location of various articulators in the vocal tract. One form of it has been made freely available in the MOCHA database [1]. This data is interesting in that it provides direct information on the production of speech, but there is the question of whether it actually provides information beyond what can be derived from the audio signal, which is much easier to collect. Although there has been some success in improving small-scale speech recognition and in demonstrating mappings between articulatory positions and spectral features of the audio signal, there are many problems to which this data has not been applied. This work investigates the possibility of using articulatory position data to improve voice transformation, which is the process of making speech from one person sound as if it had been spoken by another. After further investigation, it appears to be difficult to use articulatory position data to improve voice transformation using state-of-the-art voice transformation techniques as we only had a few positive results across a range of experiments. To achieve these results, it was necessary to modify our baseline voice transformation approach and/or consider features derived from the articulatory positions.

1. Introduction

Articulatory position data is information on the location of articulators during speech. The particular set investigated here is the freely available MOCHA database [1], which includes recordings of the 460-sentence British TIMIT corpus along with coordinates in the mid-sagittal plane for the upper and lower lip, the lower incisor, three points on the tongue, and the velum of each speaker. As this data provides direct information on the physical production of speech, there is hope that it can be used to improve models for speech. In many cases, current speech models are based on features derived from the audio signal through signal processing techniques such as LPC, cepstra, or mel-cepstral coefficients. Such features are arguably either more related to the perception of speech than the production of speech or represent an attempt to indirectly reconstruct information about production. Articulatory position data is exciting in that it gives direct information about production, but it is not without its limitations. One difficulty is that it may not fully represent the important parts of production. Seven points in a plane may not be sufficient to represent lateral effects, constrictions in the vocal tract, or the shape of the tongue. Information about pitch and power will not be directly represented. However, there may still be usable information even though the information is not complete, and there is evidence, at least for speech recognition, that it can help [2].

Another difficulty is that articulatory position data is hard to collect and this makes it fairly sparse. In most cases, it will probably not be collected during audio recordings. Thus, there is the additional question of whether this data can be useful in cases when it is available for a different speaker than the one who was recorded. There has been some work in this area as well [3]. In this context, it is natural to ask whether using articulatory position data can provide useful modeling information beyond what is available from the audio signal and for what tasks is it helpful.

Figure 1: *Transformation of Articulatory and Speech Data*



This paper attempts to extend the use of articulatory position data to voice transformation. Voice transformation is the process of making speech from one speaker sound as if it came from another. It is an important topic in speech synthesis, because successful voice transformation could greatly reduce the difficulty in producing synthetic voices with new identities and styles. Creating a concatenative speech synthesizer typically requires recording more than a thousand sentences for reasonable coverage of phonetic events. Coverage of different styles may require even more recordings. These recordings must be created for each speaker. Voice transformation has a much smaller incremental cost. After the first speaker is recorded, it is typical to record only an additional 20-30 sentences to create a new synthetic voice.

Researchers have investigated voice transformation for over 20 years and have explored many different techniques. The experiments in this paper are based on Gaussian Mixture Model (GMM) mapping techniques. These models were used at least as early as the mid-1990s [4], have been refined since then [5] [6] [7] [8], and are still considered state-of-the-art. Furthermore, scripts for implementing this type of voice transformation, based on the work of Tomoki Toda, are freely available from the FestVox website [9]. We modified these scripts to allow the use of additional features in the GMM mappings.

A high-level view of the approach taken in this paper can be seen in Figure 1. The general idea is that, in addition to mapping features derived from the speech signal data from one speaker to another, we can also map features derived from articulatory data from one speaker to another. In this paper we focus on comparing joint mappings of the speech signal and articulatory features from one speaker to another and how they compare to mappings that use only speech signal features.

2. MOCHA Database

For each of its speakers, the MOCHA database supplies audio files, Electro-Magneto Articulograph (EMA) files, laryngograph files, and electroglottograph files for the 460 sentences in the British TIMIT corpus [1]. There are two speakers for whom full data has been released. They are labeled msak0 and fsew0. The msak0 speaker is male and has a northern English accent. The fsew0 speaker is female and has a southern English accent.

The following experiments are based on features derived from the audio files and the EMA files. The audio files contain 16 bit samples at a rate of 16kHz. The EMA files contain samples at a rate of 500Hz of the x and y coordinates in the mid-sagittal plane of the positions of 7 different articulators, for a total of 14 values per sample. These 7 articulators include the upper and lower lip, the lower incisor, three points on the tongue, and the velum. The EMA files also contain additional coordinates for the bridge of the nose and the upper incisor, but they are only used for calibrating the positions of the other articulators and are not used as features in the following experiments.

3. Voice Transformation with GMM mapping

The basic idea behind GMM mapping techniques is that the probability of a joint feature vector, x, composed of features from both a source and target speaker, can be modeled by a GMM, which has the following probability density function:

$$p(x) = \sum_{i=1}^{M} \alpha_i \mathcal{N}(x; \mu_i, \Sigma_i)$$

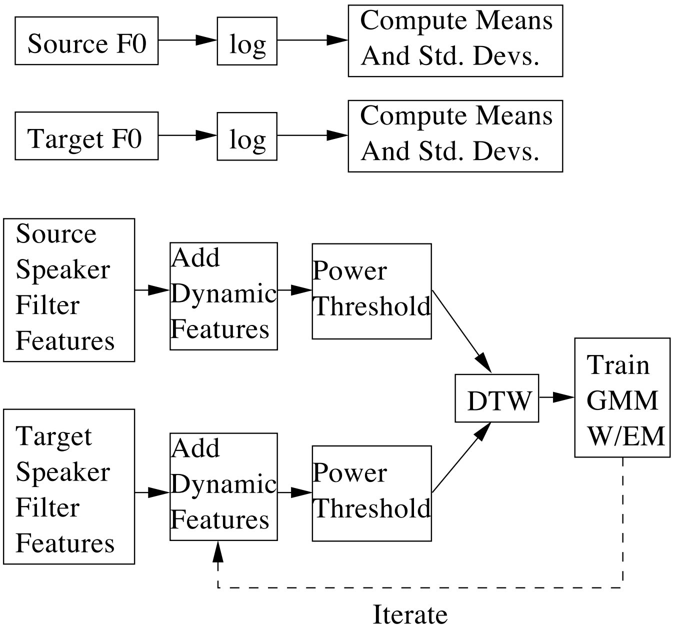
where M is the number of Gaussian components, $\mathcal{N}$ is a Gaussian distribution, μ_i and Σ_i are the mean and covariance of the ith Gaussian distribution, and the α_is are weights that are non-negative and sum to 1. In the following experiments, the default settings of the voice transformation scripts in FestVox are used to specify the form of the covariance matrix, which is diagonal in each quarter.

3.1. Training

The voice transformation training process is illustrated in Figure 2. It is based on recordings of the source and target speakers reading the same text. Fundamental frequency estimates are made for both speakers every 5ms, and mean and standard deviation statistics for their log values are calculated and recorded.

There is a separate part of the process that involves training a GMM based on filter features. The filter features used in the baseline system are the defaults used by the scripts from FestVox. 24 frequency-warped cepstral coefficients, called MCEPs, are extracted every 5ms from the recordings of the source and target speakers reading the same sentences. MCEPs approximate mel-cepstral coefficients and can be used with pitch estimates as inputs to the Mel Log Spectral Approximation (MLSA) filter [10], which is used to synthesize the transformed utterances. Dynamic features are also produced for the MCEP vectors using a weighted window centered on the current MCEP vector with values $[-0.5, 1, 0.5]$. At this point, there are now twice as many features for each speaker per frame. Frames below a certain power threshold are removed to reduce the chance of including background noise in the data. Because the durations of the parallel utterances will probably differ, dynamic time warping is used to align MCEP vectors between the two speakers to produce joint vectors with lengths of 4 times the original feature vectors (the original source speaker features plus the source speaker dynamic features plus the original target speaker features plus the target speaker dynamic features). The joint vectors are the ones that are modeled by the GMM, which is trained using EM. A couple iterations are performed where the trained GMM parameters are used to produce predictions from the source speech, which are then used to refine the DTW.

Figure 2: *Voice Transformation Training*



3.2. Transformation

Transformation is performed by the following process, which is illustrated in Figure 3:

1. Extract power, F_0, filter features (MCEP and possibly additional EMA values), and dynamic features from the utterance to be transformed.
2. Use a z-score mapping in the log domain to transform the source speaker's F_0 estimates to the target speaker's F_0 predictions.
3. Use the GMM to map the source speaker's features to the target speaker's by fixing the source speaker values and producing maximum likelihood estimates for the target speaker's features.
4. Use Maximum Likelihood Parameter Generation (MLPG) with global variance to predict final values based on filter features and dynamic features [11].
5. Use the power from the source speaker's utterance along with the F_0 and MCEP predictions as inputs to the MLSA filter to synthesize the transformed utterance.

Figure 3: *Voice Transformation*

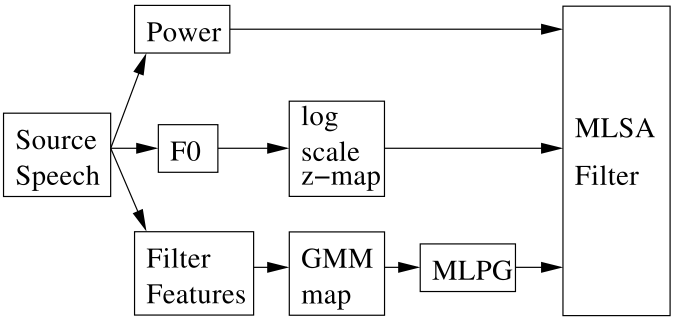


Table 1: *MCEP vs. EMAMCEP MCD Means (Std. Devs.)*

	msak0 to fsew0		fsew0 to msak0	
M	MCEP	EMAMCEP	MCEP	EMAMCEP
1	6.33(1.62)	6.88(1.61)	5.59(1.59)	5.95(1.68)
2	5.84(1.95)	6.34(1.97)	5.51(1.59)	5.79(1.71)
4	5.67(1.94)	6.25(2.06)	5.57(1.42)	5.81(1.64)
8	5.74(1.78)	6.60(1.65)	5.31(1.55)	5.95(1.62)
16	**5.58(1.79)**	6.09(1.89)	5.20(1.58)	5.46(1.62)
32	5.74(1.79)	N/A	5.06(1.62)	5.66(1.50)
64	5.74(1.70)	N/A	**5.01(1.63)**	N/A
128	N/A	N/A	N/A	N/A

3.3. Error Measure

Mel-Cepstral Distortion (MCD) is an objective error measure that is used in the following experiments to compare transformed utterances to reference utterances recorded by the target speaker. MCD has some correlation with results from subjective listening evaluations and has been used to measure the quality of voice transformation results in other work [7]. MCD is essentially a weighted Euclidean distance, that is defined by

$$MCD = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}(m_d^{(t)} - m_d^{(r)})^2}$$

where $m_d^{(t)}$ is the dth MCEP of a frame of transformed speech, and $m_d^{(r)}$ is the dth MCEP of the corresponding frame in the reference utterance recorded by the target speaker. Again, because the utterances will probably differ in length, Dynamic Time Warping is used to align them before computing the MCDs.

MCD is more related to filter characteristics of the vocal tract. Although characteristics such as power and fundamental frequency are also important to the quality of voice transformation output, the use of MCD for these experiments seems appropriate as the articulatory positions are expected to be most closely related to the filter characteristics of the vocal tract.

For the following results, no power thresholding was performed on frames before calculating MCDs, and the transformed MCEPs were used, as opposed to MCEPs rederived after synthesizing waveforms.

4. Adding Articulatory Position Data

Numerous experiments were conducted which added articulatory position data to the baseline MCEP features within the same general framework. The scripts were modified to allow the use of articulatory position features instead of or in addition to MCEP features. The rest of the processing continued in the same basic manner, with the exception that the error measure for the combination of articulatory position data and MCEPs was based solely on the MCEP subset. In the following descriptions, EMA will be used to refer to the articulatory position data, because it is the abbreviation for Electro-Magneto Articulograph, which is the specific type of articulatory position data that we used. Similarly, EMAMCEP will be used to refer to the combined use of EMA and MCEP data.

The EMA data from the MOCHA data had to be processed before combination with the MCEPs because it was sampled every 2ms instead of every 5ms, and the durations of the EMA files did not always match the durations of the audio files. Resampling was performed with the ch_track program from the Edinburgh Speech Tools [9], and EMA or MCEP features were truncated when the lengths didn't match.

Recordings from two speakers, msak0 and fsew0, were available from the MOCHA database. The experiments include transformations from each speaker to the other. The data was split into a training set of 414 utterances and a test set of 46 utterances. Most of the experiments were trained on a subset of 50 utterances due to the amount of time necessary to train the entire training set and the similarity of the results in some preliminary experiments.

Finally, there were some additional considerations that allowed the training of the GMM to work. The original EMA values were measured in thousandths of centimeters, and in some cases exceeded 5,000. Using these original values led to overflow errors with the training program, so we z-scored the EMA values to put them in a manageable range. Also, the number of Gaussian components in the GMM could affect whether training succeeded. In some cases the training program was unable to estimate parameters for the GMM and returned an error message suggesting that fewer Gaussian components should be used. In the following tables, the results for these trials will be marked as N/A (Not Applicable).

We tried to use multiple values to determine a range of success and also to track where increasing the number of components improved performance. After the initial trials, our basic choices were 16, 32, 64, or 128 components. These generally appeared to capture the range where results first improved and then worsened, presumably due to overtraining, or training even failing.

5. Experiments

5.1. Baseline Experiments

The first experiment was a comparison of only using MCEP features with using a combination of MCEP and EMA features. The only change made to the GMM mapping procedure for the initial trials including EMA was to include the EMA values in the feature vectors as well as the MCEP values. The results are in Table 1.

Adding all the EMA features directly as z-scored x and y coordinates in the mid-sagittal plane did not help in any of the trials, so it was necessary to investigate the data and the learning process more closely.

5.2. Attempts to Remove Noise from the Data

One possibility was that there was noise in the EMA data. Some potential causes were:

- The electrical apparatus originally used to collect the

Table 2: *Drift Correction MCD Means (Std. Devs.)*

	msak0 to fsew0	fsew0 to msak0
M	EMAMCEP	EMAMCEP
16	6.09(1.73)	5.58(1.59)
32	N/A	5.31(1.78)
64	N/A	N/A
128	N/A	N/A

Table 3: *First EMA Repeated MCD Means (Std. Devs.)*

	msak0 to fsew0		fsew0 to msak0	
M	MCEP	EMAMCEP	MCEP	EMAMCEP
16	5.54(1.80)	6.16(1.84)	5.18(1.59)	5.49(1.62)
32	5.67(1.82)	N/A	5.04(1.61)	5.45(1.71)
64	5.80(1.90)	N/A	5.02(1.61)	N/A
128	N/A	N/A	N/A	N/A

data

- The alignment of the MCEP with the EMA
- The resampling of the EMA data to match the default MCEP sampling rate

It has been noted by others [2] that there appears to be line noise at 50Hz in the MOCHA data. For that reason and also assuming that the motions of the articulators would be slow enough at our sampling rate, we tried applying low-pass filters with cut-offs of 45Hz and 10Hz to the MOCHA data using the sigfilter program from the Edinburgh Speech Tools [9]. Adding this low-pass filtered EMA data to the MCEP data failed to reduce the MCD error when compared to only using the MCEP data for voice transformation.

Another possible problem with the MOCHA data is that the means of the feature positions appear to vary over time more than what would be expected based on the differing phonetic contexts alone, according to other researchers [12] [13]. Although these sources were not certain whether this "drift" came from the Electo-Magneto Articulograph or the adjustment of speakers to the probes used to measure them, they found for their tasks that it was useful to try to compensate for it. We tried applying the "drift correction" strategy from the latter reference to the EMA data. This consisted of treating the mean values per utterance of the EMA features as signals, low-pass filtering these signals forward and backward with a FIR filter of length 100 and cut-off of 0.04π, and subtracting the resulting per-utterance "drift" values from the corresponding EMA features in the corresponding utterances. Adding the resulting drift-corrected data to the MCEP data failed to reduce the MCD error when compared to using the MCEP data alone for voice transformation, as can be seen in Table 2.

Another possible problem was that the EMA data was not aligned with the MCEP data. We experimented by shifting the EMA data one frame by repeating the first EMA frame. The results of this experiment are in Table 3.

This only made a minor change to the results and demonstrated that shifting the EMA by repeating the first EMA frame did not help. A companion experiment was performed where the first EMA frame was removed from each utterance. Shifting the EMA frames in that direction did not lead to an improvement in the results for trials using EMA data either. The results for this experiment are in Table 4. In both of these experiments, due to differences in the truncation of the feature files after alignment, there are small differences in the results for the trials which only used MCEP data.

Table 4: *First EMA Deleted MCD Means (Std. Devs.)*

	msak0 to fsew0		fsew0 to msak0	
M	MCEP	EMAMCEP	MCEP	EMAMCEP
16	5.54(1.80)	6.15(1.79)	5.18(1.59)	5.47(1.59)
32	**5.47(1.76)**	N/A	5.06(1.59)	5.69(1.67)
64	5.65(1.61)	N/A	**4.99(1.61)**	N/A
128	5.81(1.78)	N/A	N/A	N/A

Table 5: *DTW Based only on MCEPs MCD Means (Std. Devs.)*

	msak0 to fsew0	fsew0 to msak0
M	EMAMCEP	EMAMCEP
16	**5.84(1.81)**	5.35(1.73)
32	5.90(1.76)	**5.31(1.77)**
64	N/A	N/A
128	N/A	N/A

5.3. Attempts to Refine the Transformation Process

The baseline script that was used to perform voice transformation was based on techniques that were refined over time to handle MCEP data. It was unclear whether parts of this process were still appropriate when adding EMA data to the MCEP vectors. We investigated the following areas more closely:

- Dynamic Time Warping (DTW) used for alignment of the two speakers
- Use of the Maximum Likelihood Parameter Generation (MLPG) algorithm
- Use of multiple iterations of DTW during training

In the baseline voice transformation system, DTW was performed over all features and their derived dynamic features to align feature vectors between speakers. The distance measure used in the DTW was Euclidean. Because the MCEP and z-scored EMA values were not of the same scale, this did not seem appropriate. For this reason, we ran experiments that only considered the MCEP values during DTW when additional EMA features were used. The results are in Table 5. As can be seen through comparison with Table 1, this approach did not give better results than using MCEP data alone for the entire process. However, it did improve the results of the trials that included EMA data in comparison to previous trials that used EMA data, so it was used in later experiments.

One other thing to note is that basing the DTW only on MCEP features in the trials that also include EMA data leads to the same source speaker and target speaker frames being aligned across the different trials. This is not guaranteed when the DTW in the trials using EMA data also uses EMA values.

In the baseline voice transformation system, a program called MLPG is used to take the GMM estimates of the target speaker's MCEP and MCEP dynamic feature means and covariances to try to estimate final MCEP values that form a good path. It was unclear whether including EMA features in this process was appropriate. We ran another set of experiments where we used the means of the MCEP features for predictions and did not use MLPG (in addition to using the abovementioned strategy of only considering MCEP and MCEP dynamic feature

Table 6: *No MLPG and MCEP DTW MCD Means (Std. Devs.)*

	msak0 to fsew0		fsew0 to msak0	
M	MCEP	EMAMCEP	MCEP	EMAMCEP
16	**5.39(1.78)**	5.49(1.86)	4.95(1.57)	4.97(1.86)
32	**5.60(1.78)**	**5.50(1.81)**	4.91(1.59)	4.97(1.83)
64	5.76(1.84)	N/A	5.10(1.69)	N/A
128	N/A	N/A	N/A	N/A

Table 7: *Lip Distance MCD Means (Std. Devs.)*

	msak0 to fsew0	fsew0 to msak0
M	EMAMCEP	EMAMCEP
16	5.64(1.96)	5.40(1.78)
32	**5.55(2.00)**	5.25(1.80)
64	6.07(2.08)	5.19(1.81)
128	6.01(2.11)	5.19(1.89)

Table 8: *2-D EMA Distances MCD Means (Std. Devs.)*

	msak0 to fsew0	fsew0 to msak0
M	EMAMCEP	EMAMCEP
16	5.47(1.99)	5.21(1.73)
32	5.62(2.01)	5.14(1.80)
64	**5.56(2.02)**	N/A
128	N/A	N/A

Table 9: *EMA Projection MCD Means (Std. Devs.)*

	msak0 to fsew0	fsew0 to msak0
M	EMAMCEP	EMAMCEP
16	5.60(1.78)	5.01(1.85)
32	**5.36(1.97)**	5.00(1.86)
64	N/A	N/A
128	N/A	N/A

values during DTW). The results of these experiments are in Table 6. Adding EMA data helped in the trial that used 32 Gaussian components for the transformation from msak0 to fsew0. However, this was not a global best result for this transformation direction as the 16 Gaussian trial using only MCEP data still had better results.

5.4. Representation of EMA Features

Another possibility was that the x and y coordinates in the EMA data were a poor match for voice transformation in general or even the GMM mapping technique in particular. Perhaps there is more relevant information in features that are derived from these coordinates. After all, the x and y coordinate values are related to each other, both in terms of pairs being related to the same articulators, and in the sense that the positions of some articulators can pose constraints on the positions of others. Furthermore, the positions of some articulators relative to others provide information on constrictions in the vocal tract, which influence the filter characteristics. We investigated the following types of derived EMA features:

- Distances between the lips
- 1st order differences
- Projections onto lines fit to the articulator data

One type of vocal tract constriction that seemed reasonable to measure from the 7 articulators available in the MOCHA database was the distance between the lips. The two-dimensional Euclidean distance between the lips was used as a derived feature. The results for this experiment are in Table 7. In comparison with Table 1, it can be seen that adding lip distance improved the MCD when transforming from the msak0 voice to the fsew0 voice with 32 Gaussian components in the GMM.

Another thought was that capturing information about the motion of the articulators in two-dimensional space might supply more information. We ran experiments where the two-dimensional Euclidean distances were calculated between (x, y) coordinate pairs from frame to frame. This constructed 7 EMA derived features that could be added to the MCEP data. In this case, the dynamic features for the EMA are akin to second order differences. These trials were performed using only the MCEP and MCEP dynamic features for DTW and did not use MLPG. The results of these experiments are in Table 8. As can be seen by comparison with Table 6, adding these EMA derived distance features helped in the case of using 64 Gaussian components for the transformation from msak0 to fsew0. However, this was not a global positive result for the msak0 to fsew0 transformation as it did not perform as well as the 16 and 32 Gaussian component trials which only used MCEP data.

One problem with using 2-dimensional distances as features is that it does not include any notion of directionality, which seems like it should be important. There is a question of how to include this directionality in a meaningful way in the vectors used in the GMM mapping strategy. Although the articulator positions were measured in two dimensions, in many cases it appeared that individual articulators moved more along certain lines than others. For example, the lower incisor data showed more motion along the y-dimension than the x-dimension. In an attempt to capture some of this information, we derived features from the EMA by running linear regression on the (x, y) coordinate pairs in the training set for individual articulators to create best-fit lines, projecting the EMA (x, y) pairs onto these lines, and determining how far along these lines the articulators were. The results of using these projected EMA features are in Table 9. Again, in these trials, only the MCEP features were used for DTW and MLPG was not used. By comparison with Table 6, it can be seen that not only does adding these features improve the trial using 32 Gaussians for the transformation from msak0 to fsew0, but that this is a global positive result as it is better than all the other trials for transforming msak0 to fsew0, including the ones that only use MCEP data.

A different approach to investigating the possibility of the data being a mismatch for the model is to switch the model instead of changing the features. To this end, we tried using wagon, the Classification And Regression Tree (CART) program from the Edinburgh Speech Tools [9], instead of GMM mapping to learn the mapping between speakers. Using a step size of 100, CART predicted MCEPs from MCEPs in the fsew0 to msak0 direction with a MCD mean of 4.71 and standard deviation of 1.71. Using the combination of EMA data with MCEPs from the fsew0 speaker to predict MCEPs for the msak0 speaker gave a MCD mean of 5.22 and standard deviation of 1.90. Even with a different learning algorithm, adding EMA data failed to help improve voice transformation in terms of MCD. Although the numbers for the individual trials were better than for the GMM mapping baseline, there was the same general trend of the MCEP-only trial performing better than a trial that adds EMA x and y coordinates directly.

6. Conclusions

A number of strategies were applied to the problem of trying to use EMA data to improve a fairly standard GMM mapping based voice transformation technique in terms of Mel-Cepstral Distortion. For the most straightforward extension of the baseline voice transformation technique, none of the experimental trials that used additional EMA data directly as x and y coordinates improved the Mel-Cepstral Distortion. We made a number of attempts to use the EMA data to improve results. These attempts focused on the following three areas:

1. Removing noise from the data
2. Modifying parts of the voice transformation process that no longer appeared appropriate when using a combination of EMA and MCEP data
3. Finding a better way of representing EMA information in the model

In the first case, attempts to remove noise through filtering and realigning the EMA data, among other things, did not appear to help. In the second case, changing the way DTW was performed and not using MLPG led to results for the trials that used EMA to improve to the point where there was a trial where adding the EMA data led to better performance than using MCEP data alone. However, this was still not a global positive result as there was an MCEP trial with a different number of Gaussian components that outperformed it. In the third case, there was another positive result that came from using the distance between the lips, and finally, the first global positive result appeared in the case of using features derived from EMA by projecting the coordinates onto lines fit to the data through linear regression. In this case, the strategies of basing the DTW only on the MCEP data and not using MLPG were also followed.

It appears that the use of EMA data to improve voice transformation is not very straightforward. One additional thing to note is that all of the positive results occurred while transforming from msak0 to fsew0. There were none in the other direction. This appears to be another case of asymmetry in voice transformation. Asymmetric results have also been noted in identity perception for voice transformation [14].

There are numerous areas for further investigation. Maybe the Mel-Cepstral Distortion metric is not good enough for this task, even though it shows some correlation to subjective listening tests. Perhaps the information necessary for voice transformation is already present in MCEPs and EMA provides nothing additional. It is also possible that EMA features need to be combined or represented in a different space before they will be useful. Further experimentation will be necessary to tell.

7. 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.

8. References

[1] A. Wrench, “The MOCHA-TIMIT articulatory database,” 1999, queen Margaret University College, http://www.cstr.ed.ac.uk/artic/mocha.html.

[2] E. Uraga and T. Hain, “Automatic speech recognition experiments with articulatory data,” in *Interspeech 2006*, 2006.

[3] A. Toth, “Cross-speaker articulatory position data for phonetic feature prediction,” in *Interspeech2005*, Lisboa, Portugal, 2005.

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

[5] A. Kain, “High resolution voice transformation,” Ph.D. dissertation, OGI School of Science and Engineering, Oregon Health and Science University, 2001.

[6] T. Toda, “High-quality and flexible speech synthesis with segment selection and voice conversion,” Ph.D. dissertation, Nara Institute of Science and Technology, 2003.

[7] 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.

[8] ——, “Acoustic-to-articulatory inversion mapping with gaussian mixture model,” in *Proc. ICSLP2004*, Oct. 2004, pp. 1129–1132.

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

[10] S. Imai, “Cepstral analysis synthesis on the mel frequency scale,” in *Proceedings of ICASSP 83*, Boston, MA, 1983, pp. 93–96.

[11] T. Toda, A. Black, and K. Tokuda, “Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter,” in *Proc. ICASSP2005*, vol. 1, Philadelphia, PA, USA, Mar. 2006, pp. 9–12.

[12] K. Richmond, “Estimating articulatory parameters from the acoustic speech signal,” Ph.D. dissertation, CSTR, University of Edinburgh, 2001.

[13] Y. Shiga, “Precise estimation of vocal tract and voice source characteristics,” Ph.D. dissertation, CSTR, University of Edinburgh, 2005.

[14] A. Toth and A. Black, “Visual evaluation of voice transformation based on knowledge of speaker,” in *ICASSP*, Toulouse, France, 2006.