A Comparison of Acoustic Features for Articulatory Inversion

Chao Qin and Miguel Á. Carreira-Perpiñán

Dept. of Computer Science and Electrical Engineering
OGI School of Science and Engineering, Oregon Health and Science University
20000 NW Walker Road, Beaverton, OR 97006, USA
{cqin,miguel}@csee.ogi.edu

Abstract

We study empirically the best acoustic parameterization for articulatory inversion (the problem of recovering the sequence of vocal tract shapes that produce a given acoustic speech signal). We compare all combinations of the following factors: 1) popular acoustic features such as MFCC and PLP with and without dynamic features; 2) different short-time window lengths; 3) different levels of smoothing of the acoustic temporal trajectories. Experimental results on a real speech production database show consistent improvement when using features closely related to the vocal tract (in particular LSF), dynamic features, and large window length and smoothing (which reduce the jaggedness of the acoustic trajectory). Further improvements are obtained with a 15 ms time delay between acoustic and articulatory frames. However, the improvement attained over other combinations is very small (at most 0.3 mm RMSE).

Index Terms: acoustic-to-articulatory mapping, articulatory inversion, acoustic features, MOCHA database

1. Introduction

Articulatory inversion, or acoustic-to-articulatory mapping, consists of recovering sequences of vocal tract shapes that produce a certain acoustic signal. It has been traditionally believed to be characterized by a multi-valued mapping. That is, multiple vocal tract shapes could produce the same acoustics. A successful solution to this problem would be a major break-through in speech research and has numerous applications. For example, in speech coding, one can replace spectral parameters with slow-varying articulatory parameters. In automatic speech recognition (ASR), articulatory information could be integrated into existing acoustic based systems to further improve the recognition performance. Finally, knowledge of speech sound articulations can provide visual aids for language learning and therapy.

Articulatory inversion is a long-standing problem and remains unsolved, in particular for unvoiced sounds. The difficulty is traditionally attributed to the multi-valued nature of the inverse mapping (but see [1]). Various computational approaches have been proposed over the years. Most approaches can be broadly divided into two groups: one ignoring the multi-valued mapping e.g. analysis-by-synthesis [2] or neural networks [3]; and another addressing the multi-valued mapping, e.g. codebook look-up [4, 5], ensemble neural networks [6] or conditional modes [7]. See [8, 9] for detailed reviews.

Front-end parameterizations such as representations of articulatory and acoustic features are important to the success of all computational approaches. Thanks to the technologies such as X-ray microbeam and electromagnetic articulography (EMA), we can use as articulatory representations the measured locations of coils on different articulators such as the tongue and the lips. For acoustic representations, it is well known that acoustic features and their parameterizations such as the short-time window length are essential to ASR performance. The same stands for the inverse mapping. To our knowledge, however, there is no work about the best acoustic parameterizations for articulatory inversion. Most previous works simply chose one of the popular acoustic features used in ASR such as Mel-frequency cepstral coefficients (MFCC) [10], filter banks (FBANK) [11], line spectral frequencies (LSF) [12], and perceptual linear prediction (PLP) [7]. None of them compare laterally all the performance of these acoustic parameterizations for the inversion task.

One major problem of existing acoustic features is the jaggedness of their temporal trajectories. While this issue may not matter for ASR, it is a significant problem for articulatory inversion. This is because, while the acoustic trajectory is very jagged, the articulatory one is very smooth. Thus, when applying a mapping (e.g. a neural net) to the acoustic vectors, their output will also be jagged and cause a large error w.r.t the articulatory target. Fig. 1 shows such an example of corresponding smooth articulatory and jagged acoustic temporal trajectories.

The goals of this study are to find out the acoustic features and parameterizations that work best for the inverse mapping, and to explore ways to alleviate the jaggedness of acoustic trajectories. Besides, we also study the effect on inversion performance of a time delay between articulatory and acoustic frames.

2. Methodological setup

2.1. Popular acoustic features

In this paper, we compare most popular acoustic features that are widely used in speech/speaker recognition, speech synthesis, and speech perception. Formants are often used in early works for synthetic articulatory and acoustic data. Their temporal trajectories are very smooth since formants change slowly with time. However, formants only provide good charaterizations for vowels and they are often difficult to estimate reliably. Since we address all types of sounds in this study, we do not include formants as one of the acoustic features to be compared. We compare the following acoustic features in decreasing order of closeness to the vocal tract shape:

Linear predictive coding (LPC) It performs the short-time spectral analysis on speech frames with an all-pole filter. It provides a good approximation to the vocal tract spectral envelope for voiced speech and achieves a reasonable source-filter separation. It is less effective for unvoiced and transient regions of speech. A variant of the LPC, LSF, is often known to be

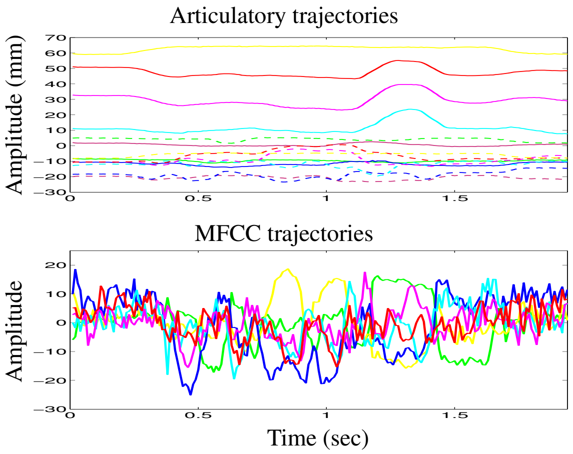

Figure 1: Smooth articulatory trajectories vs. jagged acoustic trajectories. All articulators move smoothly over time (*solid line*: horizontal locations of articulators; *dashed line*: vertical locations of articulators), while the acoustic trajectories are highly jagged over time (high-order acoustic features are more jagged than low-order ones). This means that a mapping applied to the acoustic sequence will yield a jagged articulatory sequence and thus a large inversion error. Note that for clarity, we only show odd order MFCC trajectories.

better behaved than the LPC while containing exactly the same information as the LPC [13]. Another important extension of the LPC is the LPCC, a short-time cepstral representation. The cepstral analysis is used to decorrelate dependences among variables of the acoustic features to facilitate the HMM modeling.

Filterbank analysis (FBANK) It is motivated by the fact that human ears resolve frequencies nonlinearly across the auditory spectrum. It is therefore a popular alternative to the LPC since it provides a much more straightforward way to obtain the desired nonlinear frequency resolution.

Auditory-based cepstral representations MFCC is a smoothed short-time cepstrum calculated from the log filterbank amplitude using the discrete cosine transformation. It is a robust feature containing much information about the vocal tract regardless of the source of the glottal excitation and can be used to represent all classes of speech sounds. It is the main choice for many ASR applications. Another robust variant, PLP, was originally proposed as a way of warping the spectra to minimize the differences between speakers while preserving the important speech information [14]. In addition, RASTA is a separate technique that applies a band-pass filter to the energy in each frequency subband in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel [15].

2.2. Integration of dynamic features

Instantaneous dynamic features, i.e., velocities and accelerations, are known to improve the performance of speech/speaker recognition. However, their effects on articulatory inversion remain unknown.

2.3. Variable window length

The short-time window length affects the smoothness of the acoustic features. A longer window is expected to produce smoother acoustic trajectories because two consecutive frames share more data points.

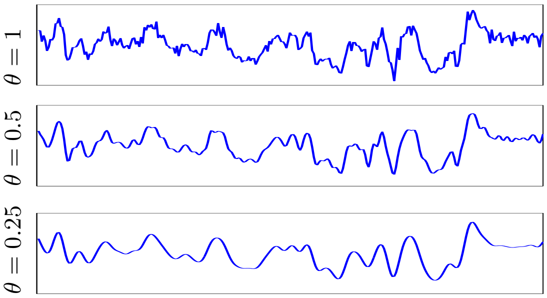

Figure 2: Illustration of smoothing of acoustic temporal trajectories. We set the cut-off frequency $\theta = 1, 0.5, 0.25$ respectively in the ascending order of smoothing levels.

2.4. Smoothing acoustic features

Smoothing the acoustic trajectories is another way to alleviate their jaggedness. It is important to keep discriminative information contained in the acoustic data as much as possible while smoothing. In this study, we perform the smoothing using a double filtering routine (implemented by `filtfilt` in Matlab Signal Processing Toolbox). First, the signal is filtered in the forward direction using an FIR filter. Second, the filtered signal is reversed and run back through the FIR filter. The cut-off frequency θ of this double filter determines the smoothing level. θ varies from 1 (no smoothing) to 0 (infinite smoothing removing all frequency components of acoustic trajectories). We use $\theta = 1, 0.5, 0.25$ respectively to denote different levels of smoothing. Fig. 2 shows an example of smoothing on a single MFCC temporal trajectory.

2.5. Articulatory inversion method

In this study, we use the neural network approach to the inverse mapping. In particular, we choose the multilayer perceptron (MLP) to map from acoustic features to articulatory ones. One appealing advantage of the neural network is that once trained, it requires much less computatational efforts compared to other methods. Although this approach ignores the multi-valued nature of the inverse mapping (since a neural net can only learn a uni-valued mapping), it is still a robust method and is useful to make fair comparisons among all sets of acoustic features. We also confirmed some of our experiments with other inversion methods (see section 3).

2.6. Performance metric

Root-mean-square error (RMSE) is one of the most commonly used performance measures for articulatory inversion. It is defined as: $e_j = \sqrt{\frac{1}{N}\sum_n \|a_j^{(n)} - b_j^{(n)}\|^2}$ where, j is articulator index, N is number of speech frames in the utterance, a and b are true and reconstructed trajectories respectively. Another popular measure is the Pearson correlation, which quantifies for a given articulator the similarity in shape between two trajectories regardless of magnitude, i.e., whether they rise and fall in synchrony: $c_j = \frac{\sum_n (a_j^{(n)} - \bar{a}_j)(b_j^{(n)} - \bar{b}_j)}{\sqrt{\sum_n (a_j^{(n)} - \bar{a}_j)^2 \sum_n (b_j^{(n)} - \bar{b}_j)^2}}$ where $\bar{a}$ and $\bar{b}$ are the means of the two sequences.

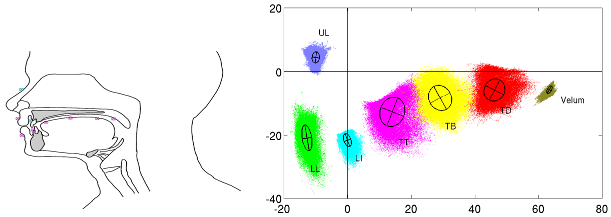

Figure 3: *Left*: pellet locations in the MOCHA database. *Right*: plot of the entire dataset for speaker `fsew0`; each pellet's data uses a different color and shows a contour line of one standard deviation centered at its mean.

3. Experiments

3.1. Experimental setup

Dataset We used the Multichannel Articulatory (MOCHA) database developed by Queen Margaret University College Edinburgh [16]. Three speakers with different regional accents have been made available so far. Four data streams are recorded synchronously for each speaker: the acoustic waves (16KHz sampling rate) together with laryngograph, electropalatograph and EMA data. The EMA recorded the movements of received coils in the midsaggital plane at 500Hz. Coils (pellets) are attached to the upper and lower lips, lower incisor, tongue tip, tongue body, tongue dorsum, and velum. Each speaker records a set of 460 British TIMIT sentences designed to be phonetically diverse. We use data from the female speaker `fsew0` with a northern English accent. Fig. 3 shows the distribution of EMA data from this speaker. We divide the dataset into two parts. One part contains 10 000 frames (randomly selected from 366 utterances) and is used for training; the other part contains 2 000 frames (from 8 unseen utterances) and is used for testing. We fix the frame shift of the short-time Hamming window to 10ms, which is roughly the average duration of speech sounds. Articulatory trajectories are downsampled to match the acoustic data. The short-time window is centered at articulatory frames.
Silence removal Exclusion of silent frames from the training set is essential to the neural network training. This is because during the silent periods, the vocal tract can in principle take any configuration. During training, given an acoustic feature vector corresponding to silence, the network would be try to map it to a large range of possible articulatory vectors, and result in a poor mapping. We apply the frame-energy based endpoint detection to remove silence, short pauses, and transient regions of speech.

Acoustic feature sets Combinations of following sets are compared: (1) Acoustic features: LPC, LSF, FBANK, MFCC, LPCC, PLP, RASTA-PLP. (2) Dynamic features: static features only and static plus dynamic features. (3) Hamming window length: 25 ms, 35 ms, 45 ms, 64 ms, 80 ms, and 96 ms. (4) Smoothing level: $\theta = 1, 0.5, 0.25$.

Inversion method We used a MLP with a single layer of 55 hidden units, trained with scaled conjugate gradients using the Netlab Toolbox for Matlab (`http://www.ncrg.aston.ac.uk/netlab`).

3.2. Comparison of acoustic features

Fig. 4 compares the performance of different acoustic feature sets. The results for RMSE and correlation are essentially equivalent. Integrating dynamic features consistently outperforms using only static features, except for PLP and RASTA-PLP. The smoother the acoustic features, the better the performance. Relatively long windows (best at 64 to 80 ms) also improve the inversion accuracy. When acoustic features are smooth enough ($\theta = 0.25$), long windows make little difference. MFCC and LPCC perform similarly. The performance of acoustic features roughly degrades in the decreasing order of closeness to the vocal tract. Among all, LSF and PLP are the best acoustic features for articulatory inversion and RASTA-PLP is (significantly) the worst. However, the overall difference is small, with all methods achieving an RMSE of 1.65 to 2 mm and a correlation of 0.56 to 0.71.

We repeated the experiment with two other methods (results not shown): using a Gaussian-mixture regression, and using a method based on representing multivalued mappings with conditional modes [7]. The results were largely the same, except that adding dynamic features did not improve as much.

3.3. Effect of time delay

Up to now, we have attempted to recover the articulatory frame at time t from the acoustic frame obtained by multiplying the acoustic wave by a short-time window centered at time t. However, it is possible that a different alignment of the articulatory and acoustic streams could result in a smaller RMSE; for example, the time spent by a wave travelling along the vocal tract may introduce a delay in the resulting acoustic waveform. While this delay would likely depend on the particular sound produced, for simplicity we consider a global delay. We conduct another empirical study to find out the best time delay, using as acoustic parameterization the best one described above (LSF with dynamic features, 64 ms window, smoothing $\theta = 0.25$). Fig. 5 shows that the best performance is achieved when the short-time window is centered 15 ms after the articulatory frame. This finding is consistent with the pilot study by Hodgen et al [17] in which a 14.4 ms time delay was found to be optimal. But again, while consistent over different experiments, the improvement over the baseline (i.e., no delay) is very small.

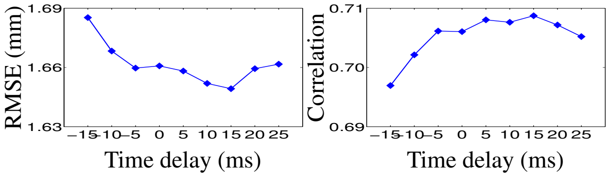

Figure 5: Effect of time delay between an acoustic frame and an articulatory frame.

4. Conclusion

We have presented an empirical study on the best acoustic parameterization for articulatory inversion. Using a simple inversion method (a neural network) as a baseline, we have compared different acoustic features, with varying lengths of the short-time Hamming window and different levels of smoothing of the acoustic temporal trajectories. Relatively large windows and smoothing were shown to alleviate the jaggedness of acoustic features. We found that best results are generally obtained with acoustic features that are more closely related to the vocal tract (in particular LSF, although PLP performed just as well), using dynamic features, 64 to 80 ms short-time window, double-filtering smoothing of cut-off frequency $\theta = 0.25$, and a 15

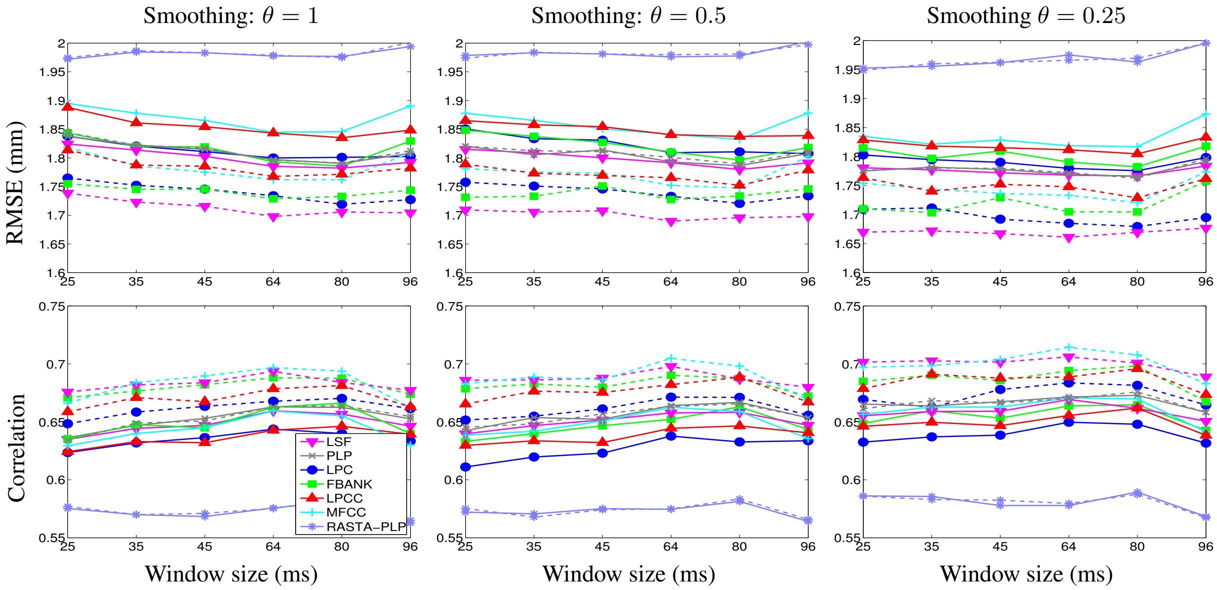

Figure 4: Performance comparison of different acoustic features (dashed/solid lines: with/without dynamic features, respectively), window size and smoothing level.

ms time delay between articulatory and acoustic frames. However, the improvement over other combinations of features or smoothing was very small (around 0.3 mm, to yield an RMSE of around 1.65 mm). These results also held when using two other inversion methods (different from the neural net).

It is important to note the limitations of our study. We used only one speaker from one database (MOCHA), and specific choices of e.g. the smoothing method. However, while other choices may alter the RMSE in absolute terms, we do not expect major changes to the relative ranking of the features (which, from fig. 4, is quite consistent over different conditions). It would also be interesting to analyze the RMSE and correlation for different categories of speech sounds.

5. Acknowledgement

Work funded by NSF CAREER award IIS–0546857.

6. References

[1] C. Qin and M. Á. Carreira-Perpiñán, "An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping," *Eurospeech 2007*, to appear.

[2] S. E. Levinson and C. E. Schmidt, "Adaptive computation of articulatory parameters from the speech signal," *J. Acoustic Soc. Amer.*, vol.74, 1983.

[3] K. Shirai and T. Kobayashi, "Estimation of articulatory motion using neural networks," *J. of Phonetics*, 1991.

[4] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *J. Acoustic Soc. Amer.*, vol. 63, 1978.

[5] J. Schroeder and M. M. Sondhi, "Dynamic programming search of articulatory codebooks," in *ICASSP'1989*.

[6] M. G. Rahim, C. C. Goodyear, W. B. Kleijn, J. Schroeter, and M. M. Sondhi, "On the use of neural networks in articulatory speech synthesis," *J. Acoustic Soc. Amer.*, vol. 93, 1993.

[7] M. Á. Carreira Perpiñán, "Reconstruction of sequential data with probablistic models and continuity constraints," in *NIPS'1999*, pp. 414–420.

[8] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal tract shapes from the speech signal," *IEEE Trans. ASSP*, vol. 2, 1994.

[9] Miguel Á. Carreira-Perpiñán, *Continuous Latent Variable Models for Dimensionality Reduction and Sequential Data Reconstruction*, Ph.D. thesis, Dept. of Computer Science, University of Sheffield, UK, 2001.

[10] S. Dusan and L. Deng, "Recovering vocal tract shapes from MFCC parameters," in *ICSLP'1998*.

[11] K. Richmond, *Estimating Articulatory Parameters from Acoustic Speech Signals*, Ph.D. thesis, University of Edinburgh, UK, 2001.

[12] S. Roweis, "Constrained hidden markov models," in *NIPS'1999*, pp. 782–788.

[13] F. Soong and B. Juang, "Line spectrum pair (LSP) and speech data compression," in *ICASSP'1984*.

[14] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoustic Soc. Amer.*, vol. 87, 1990.

[15] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. ASSP*, vol. 2, no. 4, 1994.

[16] A. A. Wrench and W. J. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in *Proc. 5th Speech Production Workshop: Models and Data*, 2000.

[17] J. Hogden, A. Löfqvist, V. Gracco, I. Zlokarnik, P. E. Rubin, and E. Saltzman, "Accurate recovery of articulator positions from acoustics: New conclusions based on human data," *J. Acoustic Soc. Amer.*, vol. 100, 1996.