On the Evaluation of Inversion Mapping Performance in the Acoustic Domain

Korin Richmond[1], *Zhenhua Ling*[2], *Junichi Yamagishi*[1], *Benigno Uría*[1]

[1]Centre for Speech Technology Research, Informatics, University of Edinburgh, Edinburgh, UK
[2]National Engineering Laboratory of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P.R.China

korin@cstr.ed.ac.uk, zhling@ustc.edu, jyamagis@inf.ed.ac.uk, benigno.uria@gmail.com

Abstract

The two measures typically used to assess the performance of an inversion mapping method, where the aim is to estimate what articulator movements gave rise to a given acoustic signal, are root mean squared (RMS) error and correlation. In this paper, we investigate whether "task-based" evaluation using an articulatory-controllable HMM-based speech synthesis system can give useful additional information to complement these measures. To assess the usefulness of this evaluation approach, we use articulator trajectories estimated by a range of different inversion mapping methods as input to the synthesiser, and measure their performance in the acoustic domain in terms of RMS error of the generated acoustic parameters and with a listening test involving 30 participants. We then compare these results with the standard RMS error and correlation measures calculated in the articulatory domain. Interestingly, in the acoustic evaluation we observe one method performs with no statistically significant difference from measured articulatory data, and cases where statistically significant differences between methods exist which are not reflected in the results of the two standard measures. From our results, we conclude such task-based evaluation can indeed provide interesting extra information, and gives a useful way to compare inversion methods.

Index Terms: Inversion mapping, evaluation, HMM synthesis

1. Introduction

Humans produce speech by moving articulators, such as the lips and tongue, to manipulate airspaces in the vocal tract, which dynamically filters and "shapes" sound energy arising from vibrating vocal folds or turbulent air movement. Manipulating articulators to produce an audible acoustic speech signal may be termed an articulatory-to-acoustic mapping. The reverse operation, taking an acoustic signal and estimating what sequence of articulator configurations might have produced it, is thus termed the *acoustic-to-articulatory mapping*. This is also known as the *inversion* mapping, since it inverts the process of speech production. More than of just theoretical interest, a reliable inversion method could help automatic speech recognition (ASR) [1], speech therapy and language training [2, 3], talking head animation and lip-syncing [4, 5, 6], and low bit-rate speech coding [7], for example.

Numerous inversion methods have been proposed, including a variety of hidden Markov models (HMM) [8, 9, 10, 11], switching linear mappings governed by a hidden Markov process [12], Kalman filtering [13], support vector regression [14], Gaussian mixture model (GMM) based regression [15], codebooks [16, 17, 18], articulatory synthesiser "mimics" [19], non-linear regression with artificial neural networks (ANN) (e.g. multilayer perceptrons (MLP) [20, 21], mixture density networks (MDN) [21], deep neural networks [22], trajectory MDNs (TMDN) [23]). In addition, several studies have looked at incorporating visual features, to give an *audiovisual-to-articulatory mapping* (e.g. [12, 14]). Alas, it has not been easy to compare studies over the years, since they have often used different data and pre-processing, though the public release of articulatory-acoustic corpora such as `MOCHA` [24, 25] and `mngu0` [26] should now reduce that problem. Despite the difficulties, though, the general trend does seem to be one of improving performance over time, according to the most often reported measures.

The two performance measures that have mainly been used are root mean square (RMS) error and correlation, calculated between each estimated articulatory trajectory and the natural, recorded one. RMS error gives an indication of the overall distance between two trajectories, while correlation indicates synchrony and similarity of shape. While these are undoubtedly useful measures, they are not necessarily ideal. First, though they can compare systems and identify the best so far, they cannot tell us when we have reached the best performance possible, or how close that *optimal inversion* is. It does not seem reasonable to reduce RMS error to zero and to obtain perfect correlation. In purely practical terms, articulography technology is imperfect, so there is unavoidably some degree of error intrinsic in the data. Moreover, significant evidence suggests multiple articulator configurations may have the same acoustic effect, so inverting this would be a one-to-many, or *ill-posed*, mapping and there may always be some "residual" error. The position of some articulators, for example, may be categorised *critical* to the production of a given phone, while others might have little or no impact on the acoustic signal [27, 20]. In addition, we cannot assume RMS error and correlation alone provide a complete and perfect performance measure. In which case, solely optimising these may not ultimately lead to optimal inversion.

With these uncertainties in mind, this paper proposes an alternative "task-based" evaluation. Several tasks could serve this purpose, for example comparing word error rates in articulation-based ASR. But, in view of the supposed many-to-one nature of the articulatory-to-acoustic mapping, it seems most compelling to evaluate inversion performance in the acoustic domain using some kind of articulatory-to-acoustic mapping. We propose to use a recently developed [28] articulatory-controlled statistical parametric synthesis system for this purpose. We find little work has previously been done on evaluating and comparing arbitrary inversion methods in the acoustic domain. Granted, articulatory synthesiser "mimics" [19] inherently optimise articulatory control parameters according to an acoustic error criterion, and acoustic error rates are often reported. But this has not been for arbitrary inversion methods; in fact, they are generally restricted to using the given articulatory synthesis model. Meanwhile, [29] used GMM-based resynthesis [15] to compare inverted articulatory parameters with natural ones. But their focus was to evaluate the feasibility of their approach to accent modification using articulation from inversion, rather than to evaluate inversion mapping performance per se. In summary, we are not aware of any previous work that has investigated whether synthesis task-based evaluation can provide additional insight into inversion performance. So, to investigate this, we evaluate four inversion methods, and broadly address two questions. First, does an acoustic evaluation provide additional information beyond RMS error and correlation scores? Second, does task-based evaluation provide any indication of how close an inversion mapping is to optimal inversion?

2. Synthesiser with articulatory control

Our experiments here use a variant of the HMM-based statistical parametric approach to speech synthesis [30] that was specifically developed to incorporate articulatory control[28]. Due to limited space, we only give a very brief overview of this *Feature-space-switched Multiple Regression HMM* (FSS-MRHMM) synthesiser, but full details may be found in [28]. In standard HMM-based synthesis, context-dependent states with distributions over segment durations, f0 and spectral features are first trained on a speech corpus. To synthesise speech, textual context features are extracted from input text which, together with state duration distributions, identify the appropriate HMM state sequence. The distributions over the spectral parameters in this state sequence may then be processed with the Maximum Output Probability Parameter Generation (MOPPG) algorithm [31] to give synthesis parameter trajectories which are sent to a vocoder (with f0 and other source parameters) to produce audible speech.

The FSS-MRHMM differs from the standard approach in several ways, the most significant of which is illustrated in Fig. 1. Instead of associating distributions over acoustic features with each state, the (spectral) distributions in the FSS-MRHMM are dependent on external input articulatory parameters. This dependency is modelled as the weighted sum of a set of linear mappings, which in turn is mediated by a *separate* "control" GMM which is fitted to the articulatory training data as an additional training step. Specifically, the distribution over acoustic features $\boldsymbol{x}_t$ given the exogenous articulatory features $\boldsymbol{y}_t$ for state $q_t = j$ at frame t is modelled as

$$b_j(\boldsymbol{x}_t|\boldsymbol{y}_t) = \sum_{k=1}^{M} \zeta_k(t)\mathcal{N}(\boldsymbol{x}_t; \boldsymbol{A}_k\boldsymbol{\xi}_t + \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \quad (1)$$

where feature vectors $\boldsymbol{x}_t \in \mathcal{R}^{3D_X}$ and $\boldsymbol{y}_t \in \mathcal{R}^{3D_Y}$ consist of static parameters and their velocity and acceleration derivatives, with static acoustic feature dimensionality D_X and articulatory dimensionality D_Y; $\mathcal{N}(;\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$; $\boldsymbol{\xi}_t = \left[\boldsymbol{y}_t^\top, 1\right]^\top \in \mathcal{R}^{3D_Y+1}$ is simply an expanded articulatory feature vector; k is the component index for the separate control GMM containing M mixture components; $\zeta_k(t) = P(m_t = k|\boldsymbol{y}_t)$ is the probability for mixture component $m_t = k$ given the articulatory features $\boldsymbol{y}_t$ at time t; and $\boldsymbol{A}_k \in \mathcal{R}^{3D_X \times 3D_Y+1}$ is the trained linear transform matrix associated with mixture component k. Hence, the acoustic distributions are made to depend on both the input articulatory feature sequence and the textual context features that govern the selection of the "residual" acoustic mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$ for each state. In order to reduce the potential for conflict and ensure a high degree of dependency upon the input articulatory features, we can modify the context features available. This is termed *context feature tailoring*. In practice, the articulatory features influence all frames, but removing context features related to vowel identity, for example, ensures vowel quality is heavily controlled by the articulatory input. In [28], it was demonstrated that by using an FSS-MRHMM with appropriate context feature tailoring it is possible to change synthesised vowel identities, or create novel vowels, just by changing the input articulatory features. Here, we propose to exploit this same articulatory control to evaluate inversion performance, as shown in Fig. 2. We test whether varying quality of articulatory trajectories, both from a range of inversion methods and natural, measured data, produce synthetic speech of equally variable quality, which can be used to differentiate performance.

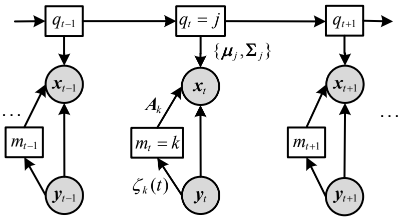

Figure 1: *The acoustic generation model in the Feature-space-switched Multiple Regression HMM [28] that we use to synthesise speech under articulatory control here.* $\mathbf{x}_t$ *is the acoustic synthesis parameter vector generated at frame t under the influence of articulatory feature vector* $\mathbf{y}_t$. *See Section 2 for details.*

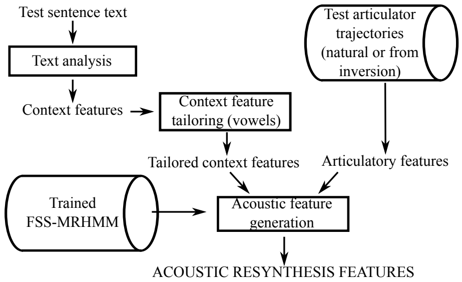
Test sentence text
Text analysis
Context features
Context feature tailoring (vowels)
Test articulator trajectories (natural or from inversion)
Tailored context features
Articulatory features
Trained FSS-MRHMM
Acoustic feature generation
ACOUSTIC RESYNTHESIS FEATURES

Figure 2: *Flowchart for synthesising test sentences under articulatory control using the FSS-MRHMM.*

3. Selected inversion mapping methods

Since our intention is to investigate whether it is useful to evaluate inversion methods in the acoustic domain using the articulatory-controlled FSS-MRHMM synthesiser, our aim is **not** to develop novel inversion methods or to challenge state-of-the-art performance. Instead, we need a *range* of methods with varying levels of performance. Accordingly, we have chosen the four inversion methods summarised below.

3.1. Linear projection

The first method is a simple linear projection from acoustic to articulatory parameters. An example of this approach to inversion is [32], who used cinefluorogram tracings of 5 speakers and linear regression to estimate midsagittal tongue shapes from formants F1-F3 during steady vowels. Although calculating the linear projection does not need an iterative optimisation algorithm, and so a validation set is unnecessary, we used the same training set as the other inversion methods for consistency. To obtain a range of inversion performance, we tried varying acoustic context window sizes: 1, 2, 4, 6, 8, and 10 acoustic frames were used.

3.2. Codebook mapping

Codebooks have been used in numerous studies, of which [16] is a well-known early example. The codebook consists of a large database of acoustic-articulatory vector pairs, either from recordings of human articulatory movements (e.g. [17]), or from sampling the parameter space of an articulatory synthesis model (e.g. [18]). To perform inversion, the database is searched to find the best vector pairs to match an input acoustic vector sequence. A variety of criteria have been tested to define what "best" means. At the simplest, Euclidean distance might be used to find the nearest acoustic vector, though there are numerous more elaborate variants of the codebook approach (e.g. [33]). The method we used is most similar to [34]. For each input vector, we find the nearest 5000 candidate acoustic vectors, using a KD-tree for efficient search. We then use Viterbi search to find the path through this sequence of candidate vector pairs that minimises the sum of a target and join cost. The target cost is the Euclidean distance between the input acoustic vector and the candidate's acoustic vector. The join cost measures the suitability of the candidate's articulatory vector for extending the paths constructed so far. Unlike [34], we use Euclidean distance between the candidate and the articulatory frame that immediately follows the path's articulatory vector in the original database. This means articulatory frames that are contiguous in the database automatically get a join cost of zero, while using the same articulatory frame at subsequent time steps in the Viterbi search does not automatically get a zero join cost. Also unlike [34], we just weighted the target and join costs equally, rather than optimising this with a validation set, since it is more important for our purposes here to have a wider *range* of inversion performance. Finally, no context window was used.

3.3. Multilayer perceptron (MLP)

The MLP is a type of ANN that is well known as a nonlinear regression method and needs little introduction here. The MLP has been used for inversion in many studies (e.g. [20, 21]). Here, we used a separate MLP for each articulator channel (i.e. each coordinate of each articulator point). Each MLP had a single hidden layer containing 100 units with a tanh activation function, and was trained using the scaled conjugate gradients (SCG) optimisation algorithm, using a validation set to decide when to stop training. A context window of 10 frames was used.

3.4. Trajectory mixture density network (TMDN)

In principle, the inversion function may feature one-to-many mappings, and the TMDN is a type of ANN that is able to take this into account by modelling full distributions over static and derived dynamic articulatory features and then using the MOPPG algorithm to generate smooth output trajectories [23]. Here, we used a TMDN with 100 units in a single hidden layer, each with a tanh activation function. We used one TMDN for each articulator channel separately, with output density functions containing 1, 2 or 4 Gaussian mixture components, and using the SCG algorithm and a validation set for training. As with the MLP, a context window of 10 frames was used.

4. Experiment

4.1. Articulatory-acoustic data

We used the electromagnetic articulography (EMA) and audio data (session 1) from the mngu0 corpus [26] for the experiments, with sensor coils attached midsagittally to the upper and lower lips, the lower incisor and the tongue tip, body and dorsum. The male British English subject was recorded reading 1263 prompts using a Carstens AG500 articulograph [35]. This records sensor coil positions in 3D Cartesian space and two angles of rotation at 200Hz sample rate. The prompts were selected from newspaper text using a Multisyn [36] text-selection tool to ensure phonetic diversity.

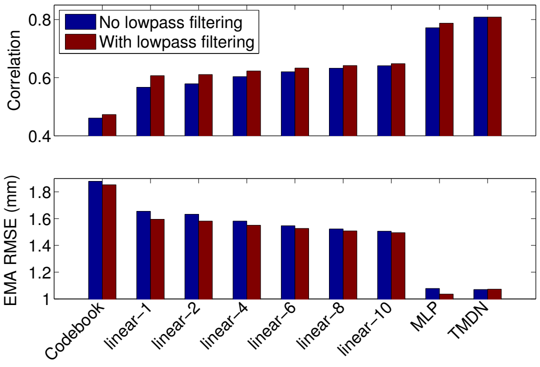

Figure 3: *RMS error and correlation for all inversion systems.*

STRAIGHT analysis [37] was used to convert the audio data to frequency-warped line spectral frequencies (LSF) of order 40, plus gain, at a frame rate to match the EMA. We used the movements of the EMA coils in the midsagittal plane only, giving a total articulatory frame size of 12 (x- and y-coordinates for 6 coils). Three subsets of the data were created: a training set of 1137 prompts without an index number ending in '0'; a validation set with 63 prompts with an odd integer preceding the final 0; and a test set with the remaining 63 prompts. Since some methods (e.g. the ANNs) are sensitive to data scaling, all EMA and LSF features vectors were z-score normalised. The data was used unnormalised to train the FSS-MRHMM synthesis system. Finally, to construct the acoustic context windows, the given number of *alternate* acoustic frames was selected, centred on the articulatory frame. So, for example, for a ten-frame context window, the time difference between the two end frames would be 90msec and the total acoustic vector size would be 410 (5msec frame shift, 41 parameters each frame).

4.2. Inversion mapping – standard measures

The inversion methods in Section 3 were trained and evaluated using the standard articulatory RMS error and correlation measures. In addition, for each inversion system we evaluated the effect of lowpass filtering using a second order Butterworth filter with 10Hz cutoff. All these results are presented in Fig. 3. The codebook gave the lowest performance, then the linear mappings, with increasing context window sizes generally improving results moderately. In all cases apart from TMDN (for which the MOPPG already provides smoothing), lowpass filtering improved results. Finally, the MLP and TMDN turned out to give similar performance in fact. This does not match previous results [23]. This may be because [23] used MOCHA, whereas here we have used mngu0. There is evidence to suggest inconsistency in the articulator positions between different sections of the MOCHA corpus [38], which is not apparent in mngu0. It is possible, therefore, that multiple Gaussian components in the TMDN gave better performance in [23] by mitigating the effects of data inconsistency, rather than by giving any benefit for the inversion mapping per se. Further investigation will be required to confirm or discount this conjecture. Overall, though, we observe a reasonable spread of performance according to these two standard measures, as desired. Finally, since we have

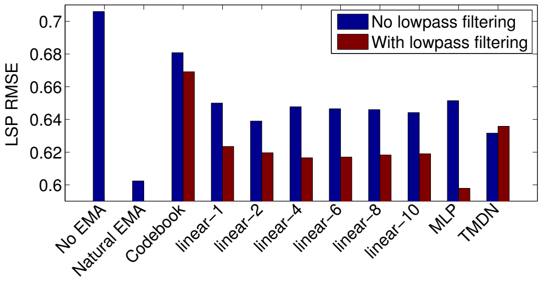

Figure 4: *Synthesised speech error for all trajectory sources.*

not sought to achieve best possible performance from each system, we stress these results should not be interpreted as an even comparison of these methods.

4.3. Synthesis evaluation

Next, we used the trajectories from all systems in Fig. 3 to synthesise the 63 sentences contained in the test set with the FSS-MRHMM system (with vowel context feature tailoring), and calculated the error of the generated acoustic features. For this we used perceptually weighted Euclidean distance, which emphasises differences where adjacent LSF coefficients are close together, corresponding to a peak in the spectral envelope (see [39], Eqs. (49–52)). These results are presented in Fig. 4. We have also included the results obtained both without EMA and with natural EMA as a bottom- and top-line comparison respectively. In many respects, these results match those in Fig. 3. As expected, having trajectories from any system is better than having no articulatory data. The ranking of the systems with no lowpass filtering is generally the same, and for all systems apart from TMDN, lowpass filtering improved results. There are some interesting differences though. The most interesting difference is that while TMDN and MLP performed similarly in terms of EMA RMS error and correlation, in terms of synthetic acoustic error, the MLP with filtering performed much better. In fact, the MLP with filtering performed on a par with natural EMA, while the TMDN appeared to perform worse than the linear systems with filtering. One explanation could be the MOPPG algorithm used in the TMDN causes over-smoothing, removing fine detail that is not reflected in the standard articulatory error measures. This needs further investigation, and will be the subject of future work. Nevertheless, this finding alone strongly motivates using synthesis, such as the FSS-MRHMM, to evaluate inversion performance in the acoustic domain.

4.4. Listening test evaluation

We conducted a listening test to verify selected observations in Fig. 4 and to gauge which differences may actually be perceived by human listeners. Thirty native British English speakers listened to 10 pairs of synthetic stimuli for each of nine preference tests in purpose-built perceptual testing booths, and were paid 5 GBP to participate. Fig. 5 indicates the nine preference tests conducted and presents the results. For the most part, these results support those in Fig. 4. The Codebook system was significantly preferred to speech synthesised without articulatory data, but was significantly worse than natural EMA. MLP, Linear10 and MDN were preferred to the Codebook system with statistical significance, but were likewise all statistically worse than natural EMA. The MLP+Filter system performed much better than both the TMDN and raw MLP output. In fact, there is no statistical difference between the MLP+Filter and natural EMA, which is a very interesting result! This is similar to [29], who reported better results with inverted articulation (after retraining) than natural recordings with a GMM-based articulatory synthesiser. This result could be interpreted as indicating *optimal inversion* has been achieved. Unfortunately, though, because of the unknown, or unquantified, inadequacies in the FSS-MRHMM synthesiser used, this would be too strong a claim. We can at least say the MLP+Filter system has achieved *sufficient* inversion performance for this task though. Finally, the preference tests indicate some of the LSF RMSE differences are imperceptible, for example there is no perceptual difference between the TMDN and Linear10+Filter output trajectories.

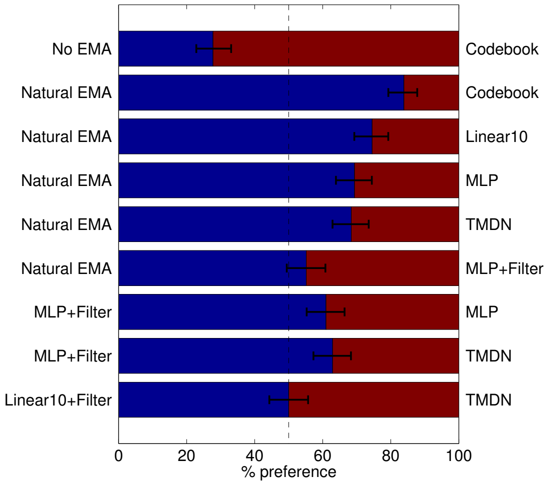

Figure 5: *Listening test results, showing the preference scores observed between the two given sets of trajectories in each subtest. The error bars indicate the 95% confidence interval.*

5. Conclusions

We have investigated the use of an articulatory-controlled HMM-based synthesiser to evaluate inversion mapping performance. There is little prior work in this direction, and our aim was simply to demonstrate whether i) such acoustic evaluation can provide useful information beyond the commonly used RMS error and correlation measures, and ii) whether such task-based evaluation can provide any indication of how close an inversion mapping is to the optimum performance possible. From our results, we conclude that this indeed *can* provide useful extra information for the purpose of comparing inversion methods. In terms of our second question, we have indeed found one inversion system performed with no statistically significant difference from the use of natural EMA trajectories. At this stage, however, we cannot conclude this system has reached optimal inversion performance, but merely that it has reached sufficient performance for the given articulatory synthesis task.

6. Acknowledgements

This work was funded in part by the National Nature Science Foundation of China (Grant No. 61273032) and the National Natural Science Foundation of China - Royal Society of Edinburgh Joint Project (Grant No. 61111130120). The research leading to these results was partly funded by the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement 256230 (LISTA), and EPSRC grants EP/I027696/1 (Ultrax) and EP/J002526/1.

7. References

[1] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 121, no. 2, pp. 723–742, 2007.

[2] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly, "Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding," *Speech Communication*, vol. 52, no. 6, pp. 493 – 503, 2010.

[3] A. B. Youssef, T. Hueber, P. Badin, and G. Bailly, "Toward a multi-speaker visual articulatory feedback system," in *Proc. Interspeech*, 2011, pp. 589–592.

[4] J. Lewis, "Automated lip-sync: Background and techniques," *The Journal of Visualization and Computer Animation*, vol. 2, no. 4, pp. 118–122, 1991.

[5] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, 1997, pp. 353–360.

[6] G. Hofer and K. Richmond, "Comparison of HMM and TMDN methods for lip synchronisation," in *Proc. Interspeech*, 2010, pp. 454–457.

[7] J. Schroeter and M. M. Sondhi, "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker Inc, 1992, ch. 8, pp. 231–268.

[8] S. Roweis, "Data driven production models for speech processing," Ph.D. dissertation, California Institute of Technology, Pasadena, California, 1999.

[9] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 175–185, 2004.

[10] L. Zhang and S. Renals, "Acoustic-articulatory modelling with the trajectory HMM," *IEEE Signal Processing Letters*, vol. 15, pp. 245–248, 2008.

[11] Z.-H. Ling, K. Richmond, and J. Yamagishi, "An analysis of HMM-based prediction of articulatory movements," *Speech Communication*, vol. 52, no. 10, pp. 834–846, 2010.

[12] A. Katsamanis, G. Papandreou, and P. Maragos, "Face active appearance modeling and speech acoustic information to recover articulation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 3, pp. 411–422, 2009.

[13] S. Dusan, "Statistical estimation of articulatory trajectories from the speech signal using dynamical and phonological constraints," Ph.D. dissertation, Dept. of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada, 2000.

[14] A. Toutios and S. Ouni, "Predicting tongue positions from acoustics and facial features," in *Proc. Interspeech*, 2011, pp. 2661–2664.

[15] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[16] 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. Acoust. Soc. Am.*, vol. 63, pp. 1535–1555, 1978.

[17] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman, "Accurate recovery of articulator positions from acoustics: New conclusions based on human data," *J. Acoust. Soc. Am.*, vol. 100, no. 3, pp. 1819–1834, 1996.

[18] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 444–460, 2005.

[19] K. Shirai and T. Kobayashi, "Estimating articulatory motion from speech wave," *Speech Communication*, vol. 5, pp. 159–170, 1986.

[20] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zachs, and S. Levy, "Inferring articulation and recognising gestures from acoustics with a neural network trained on X-ray microbeam data," *J. Acoust. Soc. Am.*, vol. 92, no. 2, pp. 688–700, 1992.

[21] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," Ph.D. dissertation, The Centre for Speech Technology Research, Edinburgh University, 2002.

[22] B. Uria, I. Murray, S. Renals, and K. Richmond, "Deep architectures for articulatory inversion," in *Proc. Interspeech*, 2012.

[23] K. Richmond, "Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion," in *Advances in Nonlinear Speech Processing, International Conference on Non-Linear Speech Processing, NOLISP 2007*, ser. Lecture Notes in Computer Science, M. Chetouani, A. Hussain, B. Gas, M. Milgram, and J.-L. Zarader, Eds., vol. 4885. Springer-Verlag Berlin Heidelberg, 2007, pp. 263–272.

[24] A. Wrench, "The MOCHA-TIMIT articulatory database," http://www.cstr.ed.ac.uk/artic/mocha.html, 1999.

[25] A. Wrench and W. J. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in *Proc. 5th Seminar on Speech Production*, 2000, pp. 305–308.

[26] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Proc. Interspeech*, 2011, pp. 1505–1508.

[27] P. J. Jackson and V. D. Singampalli, "Statistical identification of articulation constraints in the production of speech," *Speech Communication*, vol. 51, no. 8, pp. 695 – 710, 2009.

[28] Z.-H. Ling, K. Richmond, and J. Yamagishi, "Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 1, pp. 207–219, 2013.

[29] S. Aryal and R. Gutierrez-Osuna, "Articulatory inversion and synthesis: Towards articulatory-based modification of speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013.

[30] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.

[31] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp. 1315–1318.

[32] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice, "Generating vocal tract shapes from formant frequencies," *J. Acoust. Soc. Am.*, vol. 64, pp. 1027–1035, 1978.

[33] S. Demange and S. Ouni, "Acoustic-to-articulatory inversion using an episodic memory," in *Proc. ICASSP*, 2011, pp. 4620–4623.

[34] S. Suzuki, T. Okadome, and M. Honda, "Determination of articulatory positions from speech acoustics by applying dynamic articulatory constraints," in *Proc. ICSLP*, 1998, pp. 2251–2254.

[35] A. Ziert, P. Hoole, M. Honda, T. Kaburagi, and H. G. Tillman, "Extracting tongues from moving heads," in *Proc. 5th Seminar on Speech Production*, 2000, pp. 313–316.

[36] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[37] H. Kawahara, I. Masuda-Katsuse, A. de Cheveigné, and R. Patterson, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[38] K. Richmond, "Preliminary inversion mapping results with a new EMA corpus," in *Proc. Interspeech*, 2009, pp. 2835–2838.

[39] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 6, pp. 1171–1185, 2009.