Acoustic-Articulatory Modelling with the Trajectory HMM

Le Zhang, Steve Renals, *Member, IEEE,*

***Abstract*—In this letter, we introduce an HMM-based inversion system to recovery articulatory movements from speech acoustics. Trajectory HMMs are used as generative models for modelling articulatory data. Experiments on the MOCHA-TIMIT corpus indicate that the jointly trained acoustic-articulatory models are more accurate (lower RMS error) than the separately trained ones, and that trajectory HMM training results in greater accuracy compared with conventional maximum likelihood HMM training. Moreover, the system has the ability to synthesise articulatory movements directly from a textual representation.**

***Index Terms*—Trajectory HMM, Articulatory Inversion, MOCHA-TIMIT**

I. Introduction

HIDDEN Markov models (HMMs) are the standard approach to speech recognition, where the underlying task is to maximise the discrimination between similar phones or words. Speech synthesis models, on the other hand, use different techniques such as unit selection (e.g., [1]) to make the synthesised speech sound as natural as possible. This suggests that different modelling approaches may be required for recognition and synthesis; however Tokuda et al [2] have shown that the trajectory HMM formulation may be successfully applied to speech synthesis [3].

The task in which we are particularly interested is the recovery of articulatory information (the movement of human articulators) from speech acoustics, sometimes called articulatory inversion. The inversion of articulatory data involves both synthesis and recognition: we start with the acoustic signal and pose the recovery of the missing articulatory information as a synthesis problem. Conversely, the recovered articulatory information can have a complementary role in the modelling of pronunciation and acoustic variability in speech recognition.

Previous attempts to recover articulatory movement from the speech signal involved building a mapping from the acoustic domain to the articulatory domain, either manually or constructed automatically from parallel data [4], [5], [6], [7], [8], [9], [10], [11], [12]. Variations of neural networks [5], [13], [6], [11] have become popular in the latter category. Often the inversion system is built separately from the recognition framework, particularly because the slowly varying nature of articulation may be best modelled in a different way to speech acoustics which change more rapidly, and are noisier.

The authors are with the The Centre for Speech Technology Research, School of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburgh, EH8 9LW, UK. {zhang.le,s.renals}@ed.ac.uk .

Manuscript received October 17, 2007; revised December 19, 2007.

Our system, based on the trajectory HMM, differs from others in the sense that both recognition (acoustic) and synthesis (articulatory) models are constructed in the same framework, and are jointly modelled using a two-stream HMM.

The trajectory HMM extends the conventional HMM framework, and many established HMM building techniques can be reused. Moreover, in the inversion stage only the HMM state sequence is needed, so it is possible to synthesise articulator movement from a textual representation without the speech signal. We have evaluated the framework on a speaker-dependent articulatory-speech parallel corpus, MOCHA-TIMIT.

II. Trajectory HMM

Temporal derivative features, or delta features, are well known to improve the accuracy of HMM-based speech recognition systems [14]. However, the simple incorporation of delta features in an HMM leads to an inconsistent generative model [15]. These inconsistencies may be resolved by performing a per-utterance normalisation, leading to the trajectory HMM [16].

Let $\mathbf{c}$ denote the static observation vector sequence, and let $\mathbf{o}$ denote the sequence of observation vectors augmented with delta features. Then the likelihood of observing the static observation vector sequence given the HMM state sequence $\mathbf{q}$ and the model parameters λ is obtained by normalising the likelihood of obtaining the augmented observation vector sequence:

$$p(\mathbf{c} \mid \mathbf{q}, \lambda) = \frac{1}{Z_q} p(\mathbf{o} \mid \mathbf{q}, \lambda) \tag{1}$$

where Z_q is a normalisation term that depends on the state sequence:

$$Z_q = \int p(\mathbf{o} \mid \mathbf{q}, \lambda) d\mathbf{c}. \tag{2}$$

The model parameters include the Gaussian mean and variance components and can be updated using gradient-based methods.

Unlike the step-wise mean output of a conventional HMM, the mean output from $p(\mathbf{c} \mid \mathbf{q}, \lambda)$ is a smoothed trajectory, and can be used as a proper generative model, as in parametric speech synthesis. It is possible to train the trajectory HMM to maximise the generative model likelihood $p(\mathbf{c} \mid \mathbf{q}, \lambda)$. This has considerably higher complexity than conventional maximum likelihood training for HMMs, and is rarely done for HMM-based speech synthesis systems [3].

III. Acoustic-Articulatory Modelling for Articulatory Inversion

Our system starts with parallel articulatory-speech data where the movement of articulators has been recorded using an electromagnetic articulography (EMA) machine. The challenge of the task is that different articulator configurations (vocal-tract shapes) can produce the same sound, which means the mapping from speech to articulatory domain is not unique [4], [17], and that the acoustic signal is less smooth and varies faster, compared with articulator movements.

Instead of seeking a direct mapping between the acoustic and articulatory signals, our methodology centres around the idea of jointly optimising a single model for acoustic and articulatory information. The model has two parts, both using the same multi-state phone-level HMMs: an articulatory synthesis model which (given an HMM state sequence) generates a smoothed mean trajectory (1); and an alignment model which derives the state sequence for synthesis from an unseen utterance. We carried out training on the parallel data by creating a two-stream HMM where one stream is modelled by an articulatory HMM with single Gaussian output densities, and the other is the standard Gaussian mixture acoustic HMM. After that, the parameters of the articulatory stream are updated using trajectory HMM maximum likelihood estimation [16]. In this paper we choose to update Gaussian mean components only, as updating Gaussian variances was found to be both time-consuming and less effective.

For inversion we first derive a representative HMM state alignment from the acoustic channel. Then the parameter generation algorithm [2] is executed to produce the smoothed mean trajectory from (1) in the articulatory domain. One feature of the system is the flexibility in obtaining the HMM state alignment at the inversion stage. Depending on the available resources, it can be the state sequence returned by an HMM decoder, the forced alignment derived from phone labels, or the synthesised state sequence from a textual representation, using a suitable duration model. An overview of the acoustic-articulatory model is illustrated in figure 1.

Delta features, which play a central role in trajectory HMM systems, are obtained from the regression coefficients that represent the temporal slope of each feature [14]. In the HTK system[1], delta coefficients are computed from the previous and next two frames. The delta-delta coefficients are computed in the same way using the previous and next two deltas, meaning that the whole window covers nine frames. In the HTS HMM-based speech synthesis system[2] a simpler three frame window is employed, using a quadratic regression for delta-deltas. We experimented with both kinds of windows, and found that choice of window has an impact on the articulatory inversion task. We will refer to the three-frame dynamic window (as used in HTS) as *dw3*, and the nine-frame window (as used in HTK) one as *dw9*.

[1]http://htk.eng.cam.ac.uk/
[2]http://hts.sp.nitech.ac.jp/

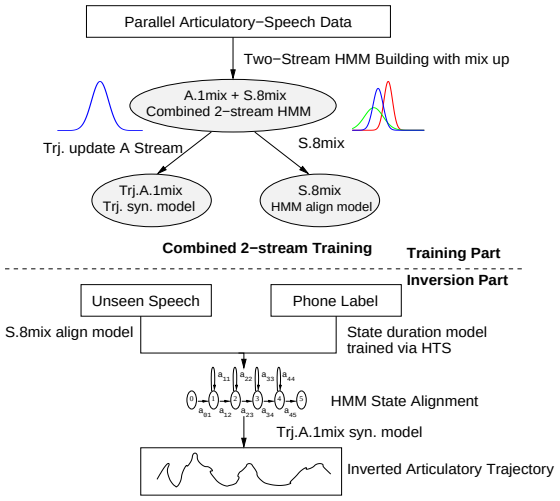

Fig. 1. *Overview of the articulatory-acoustic modelling system. Using two-stream combined training results in greater accuracy compared with the separately trained ones.*

IV. Experiments

The MOCHA-TIMIT corpus[3] is a speaker-dependent recording of TIMIT sentences with articulatory information captured using EMA, along with the acoustic signal. It includes one male speaker (msak0) and one female speaker (fsew0), each uttering 460 TIMIT sentences. Electromagnetic receiver coils are attached to 7 articulators in both x and y-coordinates during recording, providing a total 14 channels of articulatory information sampled at 500 Hz. The female data (fsew0) is used in this paper.

In preparing the experiment, we down-sampled the EMA data to 100 Hz to match the 10 ms frame-rate of the acoustic features, which are the usual 12th-order MFCCs with log-energy plus their delta and delta-deltas. All delta features are computed using the three-frame *dw3* window unless mentioned otherwise. A mean-filtering normalisation is performed to compensate some EMA measure errors introduced in the recording stage [18]. We set aside the utterances whose recording number ends with 2 for validation (46 utterances), those ending with 6 for test (46 utterances) and the remaining 368 utterances for training. The phone set consists of 45 phones including silence. The inversion performance will be reported as average RMS (root mean square) error compared with the recorded articulatory data.

Similar to building an HMM-based speech recognition system, we refined our inversion models incrementally. Both the (articulatory) synthesis model and the (acoustic) alignment model started from a single component Gaussian, three-state, left-to-right monophone model trained using HTK. Depending on the training scheme used, three synthesis models were built:

- hmm.A: baseline HMM trained on the articulatory data only, using HTK.
- trj.A: trajectory HMM, built from the baseline HMM, with the Gaussian mean components updated using the forced alignment provided by hmm.A.

[3]http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

- trj.C: jointly trained two-stream trajectory HMM, with updated Gaussian mean components for the articulatory stream. A default stream weight of 1.0 is used for both streams.

Since a trajectory HMM equivalent of the Baum-Welch algorithm has not been discovered, it is prohibitively expensive to estimate trajectory HMMs with multi-component Gaussian mixture densities. Thus our articulatory synthesis component is limited to single component Gaussian densities (1mix). In deriving the alignment for training, however, there is no such restriction. We can therefore get more accurate alignment by using more mixture components. In our experiments, Gaussian mixture densities with four (4mix) and eight (8mix) components were used to derive the HMM alignment.

To carry out the inversion, again an HMM alignment is required for each testing utterance. Different strategies can be employed to maximise resource usage. For utterances with only acoustic data, we choose to use the alignment returned by decoding speech directly using a phoneloop grammar (S-Decode). If we have access to phone labels then a better alignment can be obtained by running the decoder in forced alignment mode (S-Align). In addition, we give the result based on the forced alignment of the recorded articulatory data (A-Align). Although this information will not be available in real inversion tasks, it nevertheless gives us an indication of the "topline" performance using single Gaussian densities in the articulatory synthesis model.

A unique feature of the inversion system is the ability to perform synthesis with only phone label information. The required HMM state alignment for synthesis can be constructed from a state duration model. Using the HTS system, we built a monophone duration model from the training data. A state sequence was then synthesised using the duration model and the provided test labels and the mean trajectory was then generated.

The inversion results on test data in terms of average RMS error (mm) over the 14 channels are presented in Table I. The RMS error obtained when synthesising from the phone labels is shown in Table I as trj.dur. We also list results obtained using the same data set employing a multi-layer perceptron (MLP) [11] and a trajectory mixture-density network (TMDN) [19]. We conducted paired one-tail t-tests between the results obtained using the same number of mixture components and decoding/alignment approach (i.e. within each column in Table I). The differences between the obtained results are all significant at the $p < 0.05$ level, except where marked with †.

Compared to the result from baseline model hmm.A, the two trajectory models (trj.A and trj.C), achieve a significantly lower RMS error in the different inversion configurations. This demonstrates the effectiveness of trajectory training. Moreover, in table I we find that the jointly trained model (trj.C) results in significantly lower RMS errors than trj.A, in which the articulatory stream and speech stream are trained separately. Hence training a model jointly on the acoustic and articulatory streams results in a reduced RMS error. Furthermore, increasing the number of mixture components in the acoustic alignment model consistently reduces the RMS error, despite the fact that the final synthesis stage uses a single Gaussian model.

TABLE I
RMS ERROR (MM) OF ARTICULATORY INVERSION ON TEST DATA

Model	S-Decode			S-Align			A-Align
	1mix	4mix	8mix	1mix	4mix	8mix	1mix
hmm.A	1.936	1.901	1.876	1.842	1.814	1.804	1.679
trj.A	1.923†	1.811	1.756	1.715	1.656	1.624	1.386
trj.C	1.887	1.756	**1.705**	1.630	1.633†	**1.580**	1.477
trj.dur	2.339						
MLP	1.62						
TMDN	1.40						

Comparing the different alignment methods, it can be seen that the method based on forced alignment with phone labels (S-Align) results in significantly ($p < 0.05$) lower errors than the alignment obtained from direct decoding (S-Decode). The final column of Table I gives an upper bound on performance using a single Gaussian monophone model aligning to the recorded articulatory data (A-Align). The fact that trj.C performs worse than trj.A in this condition is because when the actual articulatory data is provided, the addition of acoustic information lowers the alignment accuracy.

The recovered trajectory for the movement of upper lip in the x direction for the first utterance in the test set is displayed in Figure 2, where the trained trajectory HMM (trj.dw3) is observed to give a better fit to the data than the baseline HMM (hmm.dw3).

We also investigated the effect of the delta coefficient regression window for this task. Using the *dw3* window for delta coefficient estimation, the best inversion result is an RMS error of 1.876 mm for the baseline HMM, and 1.580/1.705 mm for a trained trajectory HMM using an alignment derived from the S-Align/S-Decode conditions respectively, both employing an 8-component mixture model for alignment. Although not shown here, the *dw9* window results in slightly lower RMS errors than *dw3*. Among 21 results in the first three rows of Table I, we find only 8 cases where the difference between *dw3* and *dw9* is statistically significant at the 0.025 level of a two-tail paired t-test. And in only 2 instances does *dw9* have a lower RMS error than *dw3*. Figure 2 shows the recovered trajectories using the two windows, and it is clear that the *dw3* window results in a smoother estimated trajectory, compared to *dw9*.

The lowest inversion error from the the speech signal alone is 1.705 mm, which compares well with an error of 1.62 mm obtained when using an MLP for direct acoustic-articulatory mapping [11], especially since in this approach the articulatory trajectory is generated using single Gaussian densities. More recently, the TMDN approach [19] has resulted in a decreased RMS error of 1.40 mm on this data set.

V. DISCUSSION

Recent interest in the use of HMM-based systems for speech synthesis, and the development of the trajectory HMM, has resulted in a resurgence of interest in the development of unified models for speech recognition and synthesis with a principled statistical basis. In this work we use a common generative model for acoustic-articulatory data that—with appropriate marginalisation—can be used for both recognition and

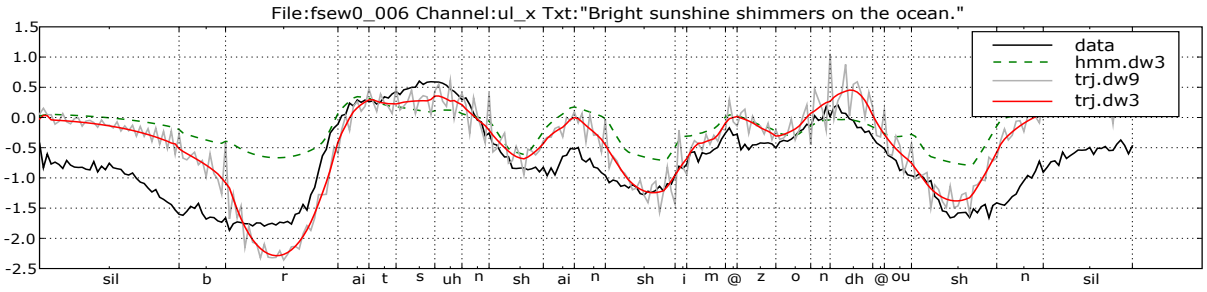

Fig. 2. *Recovered trajectory for the movement of Upper Lip in x coordinate (ul_x) of test utterance fsew0_006. The trained trajectory HMM (`trj.dw3`) shows a closer fit to the data than the baseline HMM (`hmm.dw3`), with state alignment derived from a 8-mixture jointly trained 2-stream HMM. The light gray trajectory of (`trj.dw9`) shows the noisy effect of using 9-frame dynamic window.*

synthesis of acoustic and articulatory signals. Our experiments in this paper confirm that training such a model jointly (trj.C) results in more accurate generation of articulatory trajectories, compared with separately trained models (trj.A).

Despite its theoretical attractions, the trajectory HMM has a major limitation at the current time. In the absence of a "trajectory HMM Baum-Welch" algorithm, training models with multiple component mixtures is prohibitively expensive. Thus, in this work, the articulatory synthesis model was limited to trajectory HMMs with single Gaussian densities. In the HTS speech synthesis system, this limitation is implicitly addressed through the use of detailed context. In this work we have used monophone models, and it is clear that the use of context-dependent models is worth investigating.

Although there are significant technical challenges related to trajectory HMM training, there are several advantages to pursuing the trajectory HMM as a unified model for synthesis and recognition. The fact that existing software frameworks for HMMs may be reused provides a platform for experimentation, and a principled, efficient way to initialise models (using conventional HMM parameter estimation). In the articulatory-acoustic modelling case, the use of duration modelling approaches developed in HMM-based speech synthesis enables articulatory movement to be generated without the need for acoustics, and it is also possible to apply speaker adaptation approaches used successfully in recognition and synthesis.

ACKNOWLEDGEMENTS

The authors would like to thank Korin Richmond and Junichi Yamagishi for their comments on this paper. Korin Richmond also provided the Matlab normalisation script for compensating the EMA measure errors.

REFERENCES

[1] 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.

[2] K. Tokuda, T. Yoshimura, T. K. Takashi Masuko, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. of ICASSP 2000*, Istanbul, Turkey, June 2000, pp. 1315–1318.

[3] H. Zen and T. Toda, "An overview of Nitech HMM-based speech synthesis system for Blizzard challenge 2005," in *Proc. of Interspeech 2005*, Portugal, Lisbon, September 2005, pp. 93–96.

[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," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1535–1555, May 1978.

[5] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zacks, and S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," *Journal of the Acoustical Society of America*, vol. 92, no. 2, pp. 688–700, August 1992.

[6] T. Kobayashi, M. Yagyu, and K. Shirai, "Application of neural networks to articulatory motion estimation," in *Proc. ICASSP-91*, April 1991, pp. 489–492.

[7] C. S. Blackburn and S. Young, "A self-learning predictive model of articulator movements during speech production," *Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1659–1670, March 1999.

[8] H. B. Richards, J. S. Mason, M. J. Hunt, and J. S. Bridle, "Deriving articulatory representations from speech with various excitation modes," in *Proc. ICSLP '96*, vol. 2, Philadelphia, PA, 1996, pp. 1233–1236.

[9] S. King and A. Wrench, "Dynamical system modelling of articulator movement," in *Proc. ICPhS 99*, San Francisco, Aug. 1999, pp. 2259–2262.

[10] S. Dusan and L. Deng, "Acoustic-to-articulatory inversion using dynamical and phonological constraints," in *Proceedings of the 5th Seminar on Speech Production: Models and Data*, Kloster Seeon, Germany, May 2000, pp. 237–240.

[11] K. Richmond, S. King, and P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Computer Speech and Language*, vol. 17, pp. 153–172, 2003.

[12] T. Toda, A. W. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with gaussian mixture model," in *ICSLP2004*, Jeju, Korea, 2004, pp. 1129–1132.

[13] M. G. Rahim, W. B. Keijn, J. Schroeter, and C. C. Goodyear, "Acoustic to articulatory parameter mapping using an assembly of neural networks," in *Proc. ICASSP-91*, April 1991, pp. 485–488.

[14] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. ASSP*, vol. 34, no. 1, pp. 52–59, 1986.

[15] J. Bridle, "Towards better understanding of the model implied by the use of dynamic features in HMMs," in *ICSLP 2004*, 2004, pp. 725–728.

[16] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Computer Speech & Language*, vol. 21, no. 1, pp. 153–173, January 2007.

[17] C. Qin and M. A. Carreira-Perpiñán, "An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping," in *Proceedings of INTERSPEECH 2007*, 2007, pp. 74–77.

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

[19] ——, "Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion," *In Proc. NOLISP 2007 (In Press)*, 2007.