MODELS OF SPEECH DYNAMICS IN A SEGMENTAL-HMM RECOGNIZER USING INTERMEDIATE LINEAR REPRESENTATIONS

Philip J.B. Jackson and Martin J. Russell

Sch. Electron. Elec. & Comp. Eng., University of Birmingham, UK [p.jackson@bham.ac.uk].

ABSTRACT

A theoretical and experimental analysis of a simple multi-level segmental HMM is presented in which the relationship between symbolic (phonetic) and surface (acoustic) representations of speech is regulated by an intermediate (articulatory) layer, where speech dynamics are modeled using linear trajectories. Three formant-based parameterizations and measured articulatory positions are considered as intermediate representations, from the TIMIT and MOCHA corpora respectively. The articulatory-to-acoustic mapping was performed by between 1 and 49 linear transformations. Results of phone-classification experiments demonstrate that, by appropriate choice of intermediate parameterization and mappings, it is possible to achieve close to optimal performance.

1. INTRODUCTION

The Balthasar project seeks efficient, complete and trainable models for speech processing that represent its production, and can thus characterize the mechanisms that give rise to variablity in speech. In principle, such models will accommodate production strategies used in different speaking styles, offer improved performance in adverse environments, and ultimately provide a unified framework which can support a range of speech technologies, from recognition to synthesis.

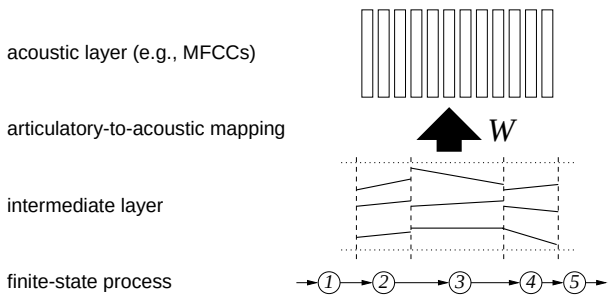

Fig. 1. Illustration of a segmental model using linear trajectories in the intermediate space and mapping function W.

The Balthasar project is funded by EPSRC grant M87146, see http://web.bham.ac.uk/p.jackson/balthasar/ for details. The authors would like to acknowledge Nick Wilkinson's help in preparing the formant data and synthesizer control parameters from TIMIT.

Acoustic features of speech, typically derived from short-term log power spectra, reflect articulatory dynamics indirectly, often as movement across frequency bands. Automatic extraction of articulatory features from acoustic data is a significant pattern processing problem and prone to error, so simple substitution with more-appropriate articulatory features is not viable. However, in the present work, states of the underlying Markov process (at symbolic/phonetic level) are associated with trajectories in an articulatory-based feature space (intermediate layer), which are mapped onto the surface (acoustic) feature space, where comparison is made with observations (see Fig. 1).

Here, we consider a simple class of Multi-level Segmental HMM (MSHMM) whose trajectories in the articulatory-based representation are linear, and whose articulatory-to-acoustic mapping is realized as a set of one or more linear mappings [1]. Non-linear mappings, such as multi-layered perceptrons and radial-basis function networks, have been investigated [2], and many kinds of trajectory tried in the acoustic domain, e.g., [3, 4, 5]. In the present case, the trajectories are also linear in the acoustic-feature space, although modifying either the intermediate representation or the mapping function would remove this property. Thus, the performance of an appropriate probabilistic linear-trajectory SHMM of the type described in [6] provides a theoretical upper bound on the performance of this type of MSHMM, and has been shown to outperform a conventional HMM [6]. Therefore, the goal of this paper is to determine whether the upper bound can be achieved by appropriate choice of articulatory representation and linear articulatory-to-acoustic mappings; it is demonstrated to be so.

2. THEORY

2.1. Linear-trajectory segment models

In the terminology of [6], the model that concerns us is a fixed, linear trajectory segmental HMM (FT-SHMM). Each state s_i of such a model is identified with a midpoint vector $\mathbf{c}_i$ and slope vector $\mathbf{m}_i$, whose dimension N is that of the acoustic-feature space. A trajectory of duration τ is defined by $\mathbf{f}_i(t) = (t - \bar{t})\,\mathbf{m}_i + \mathbf{c}_i$, where $\bar{t} = (\tau + 1)/2$, and, given

the state s_i, the probability of the sequence of acoustic vectors $\mathbf{y}_1^\tau = \{\mathbf{y}(1), \mathbf{y}(2), \ldots, \mathbf{y}(\tau)\}$ is

$$b_i(\mathbf{y}_1^\tau) = \prod_{t=1}^{\tau} \mathcal{N}\left(\mathbf{y}(t); \mathbf{f}_i(t), R_i\right), \quad (1)$$

where $\mathcal{N}(\cdot)$ denotes the multivariate Gaussian pdf with mean $\mathbf{f}_i(t)$ and diagonal $N \times N$ covariance R_i, evaluated at $\mathbf{y}(t)$. The case $\mathbf{m}_i = \mathbf{0}$ corresponds to a constant-trajectory SHMM, which is functionally identical to a conventional HMM except for an upper bound $\tau_{\max}$ on state duration.

Now consider a trajectory in the M dimensional articulatory space, $\mathbf{f}_i$, that is projected onto the acoustic layer by the mapping W, which is assumed to be linear. The probability of the acoustic sequence becomes

$$b_i(\mathbf{y}_1^\tau) = \prod_{t=1}^{\tau} \mathcal{N}\left(\mathbf{y}(t); W(\mathbf{f}_i(t)), R_i\right). \quad (2)$$

2.2. Model parameter estimation

It is assumed that simultaneous articulatory-acoustic data are available for learning the mapping. It is not necessary, as W could be optimized in Baum-Welch style re-estimation, along with the other model parameters, however, using matched data preserves the strict articulatory interpretation of the models in the intermediate layer. Given matched sequences $\mathbf{x}_1^T$ and $\mathbf{y}_1^T$ of articulatory and acoustic features, respectively, we would like to find the $N \times M$ matrix W that minimizes

$$E = \sum_{t=1}^{T} \left(W\mathbf{x}(t) - \mathbf{y}(t)\right)' R_i^{-1} \left(W\mathbf{x}(t) - \mathbf{y}(t)\right), \quad (3)$$

which is a least-squares problem, and soluble, e.g., by singular value decomposition. In general, the matched data may be partitioned into K phone categories, and a separate mapping W_k is estimated for each one $k = 1, \ldots, K$.

Let $\mathcal{M}$ be an S-state phone-level MSHMM. Once the mappings W_k have been determined, a simple extension of the segmental Viterbi decoder (see, for example [6]) can be used to compute the state sequence $\hat{\mathbf{s}}$ that maximizes

$$\Pr(\mathbf{y}, \mathbf{s}|\mathcal{M}) = \prod_{i=1}^{S} b_i\left(\mathbf{y}_{t_i}^{t_{(i+1)}-1}\right), \quad (4)$$

where t_i is the time at which the state sequence $\mathbf{s}$ enters state s_i (for simplicity, it is assumed that the probability of a transition from s_i to state s_j is zero unless $j = i + 1$).

Given $\hat{\mathbf{s}}$, it can be shown that the maximum-likelihood estimates of the midpoint $\hat{\mathbf{c}}_i$ and slope $\hat{\mathbf{m}}_i$ for state s_i are

$$\hat{\mathbf{c}}_i = \frac{1}{T} \sum_{t=t_i}^{t_{i+1}-1} \left(D_i W_k\right)^\dagger D_i \mathbf{y}(t) \quad (5)$$

$$\hat{\mathbf{m}}_i = \frac{\sum_{t=t_i}^{t_{i+1}-1} (t - \bar{t}) \left(D_i W_k\right)^\dagger D_i \mathbf{y}(t)}{\sum_{t=t_i}^{t_{i+1}-1} (t - \bar{t})^2}, \quad (6)$$

respectively, where † denotes the pseudo inverse, $D_i = R_i^{-\frac{1}{2}}$, $\bar{t} = (t_{i+1} + t_i - 1)/2$, and k is the phone category for model $\mathcal{M}$. Note that if $N = M$ and the rank of W_k is N, then $(D_i W_k)^\dagger = W_k^\dagger D_i^\dagger$ and the D_i terms disappear from both equations. Interpreting equations 5 and 6, the optimal midpoint and slope in the articulatory domain give the best linear fit to the (pseudo)inverse-transformed observation vectors, accounting for the covariance information in the matrix D_i.

3. METHOD

3.1. Experimental procedure

Two types of parameterization were considered for the intermediate layer: formant-based parameters, which were derived automatically from acoustic data in the TIMIT corpus using the Holmes formant analyzer [7], and measured articulatory parameters from the MOCHA corpus [8]. For both, linear articulatory-to-acoustic mappings were estimated with matched sequences of articulatory and acoustic data. Given these mappings, Viterbi alignment and Eqs. 5 and 6 were used to re-estimate the MSHMM state parameters. The maximum state duration was set to 10 frames ($\tau_{\max} = 10$).

With one mapping per category, a series of W_k was obtained for each categorization of the phones (number of mappings): A. all data (1); B. speech, silence/non-speech (2); C. linguistic categories (6); D. as in Deng and Ma [4] (10); E. discrete articulatory regions [2] (10); F. individual phones (49). No language model was used for the phoneme classification experiments, which used models of the normal 49 phones, and 39 phones, plus silence, for scoring.

3.2. TIMIT data

Speech from all male subjects in the TIMIT training and test sets was downsampled to 8kHz (8 dialect regions of N. American English). Acoustic features (13 MFCCs including zeroth) were obtained using HTK (25 ms window, 10 ms fixed frame rate), while formant-based parameters were extracted using the Holmes formant analyzer [7]. Three formant-based parameterizations were considered: (a) 3 formant frequencies F1, F2 and F3 (25 Hz resolution); (b) 3 formant frequencies plus 5 frequency-band energies; (c) the 12 control parameters from Holmes-Mattingly-Shearme parallel formant synthesizer, which include the amplitudes and frequencies of F1, F2 and F3. A bias input (set equal to 1) was added to all of them to allow an offset to be learnt, for each acoustic feature. Thus the dimensions of the intermediate space for (a), (b) and (c) were M = 3, 8 and 12 respectively. For (a) and (b), the analyzer also returned a confidence measure for each formant estimate, based on the curvature and relative amplitude of a candidate formant peak.

Map.	Base	(a) F1-3	(b) F1-3 + BE5	(c) PFS12	(a)* F1-3
ID_0	52.9	-	-	-	-
ID_1	54.3	-	-	-	-
A (1)	-	47.7	53.1	54.0	47.7
B (2)	-	47.5	53.2	53.9	-
C (6)	-	48.9	53.3	53.7	48.6
D (10)	-	48.9	52.9	53.5	48.7
E (10)	-	49.1	53.2	52.7	49.3
F (49)	-	52.9	53.9	54.1	53.1

Table 1. Classification accuracy (%) for TIMIT tests: ID_0, identity mapping with constant trajectory; ID_1, identity mapping with linear trajectory; A, linear mapping with linear trajectory; B, two lin. map. & lin. traj.; C, six lin. map. & lin. traj.; D, ten lin. map. & lin. traj.; E, ten lin. map. & lin. traj.; F, lin. map. per phone & lin. traj. The parameterizations were: MFCCs for the Baseline; (a) F1-3, formant frequencies; (b) F1-3 + BE5, formants and band energies; (c) PFS12, synthesis control parameters. *Maximum-likelihood estimation of the trajectory parameters.

Threshold	0.0	0.2	0.4	0.6	0.8
A (1)	53.1	53.1	53.0	52.8	52.3
F (49)	53.9	53.5	53.4	52.7	52.7

Table 2. Classification accuracy (%) for TIMIT tests varying the threshold for the minimum acceptable confidence, using formants and band energies with one mapping and 49. Key as for Table 1.

Mapping	1	2	3	4	5	Avrg.
ID_0	53.4	53.2	53.8	54.4	54.8	53.9
ID_1	55.4	55.1	55.5	56.0	56.1	55.6
A (1)	54.4	53.7	54.0	55.0	55.2	54.5
B (2)	55.3	54.0	53.9	55.0	55.1	54.7
C (6)	54.8	54.6	55.1	55.9	55.5	55.2
D (10)	54.9	54.9	55.2	55.4	55.4	55.1
E (10)	54.9	54.6	55.0	55.7	55.5	55.1
F (49)	55.1	55.2	55.0	56.1	55.8	55.4

Table 3. Classification accuracy (%) for MOCHA experiments across the five jack-knife test sets with the average. Key to mappings as for Table 1.

3.3. MOCHA data

MOCHA speech files (16 kHz), of an adult male speaker of S. British English (R.P.), were similarly pre-processed to 13 MFCCs. A principal component analysis was made of the 14 articulatory position measurements (referenced x- and y-coordinates of 7 EMA coils, 25 ms window, 10 ms frame rate). The log-energy of the laryngograph signal was appended to the first nine articulatory modes, to make a 10-coefficient feature vector. The relatively small database was divided in five ways for training, withholding every fifth file for testing, and the results averaged, as in [8].

4. RESULTS

4.1. TIMIT experiments

Baseline performance. Baseline phone classification experiments were conducted using FT-SHMMs (no intermediate articulatory layer) [6]. The results are presented in the first column of Table 1 for constant (ID_0) and linear (ID_1) trajectory FT-SHMMs, and are consistent with [6]. Thus the performance upper bound for the MSHMM experiments is 54.3% phones correct.

Multi-level segmental HMM results. The following entries in Table 1 are phone classification results for different combinations of the number of mappings and the type of formant-based intermediate layer. In these experiments, the D_i term in Eqs. 5 and 6 was ignored, so that the mappings W_k were optimized according to the minimum mean squared error criterion. In general, improved results are obtained by either increasing the dimension of the intermediate layer (column 3), or increasing the number of mappings (final row). In particular, by using the PFS12 representation or a large set (49) of mappings, near optimal performance is achieved.

Effect of formant-frequency confidence. Since optimization of the mappings W_k will be compromised by errors in the formant parameter estimates, experiments were conducted in which formant and MFCC vectors were only used in estimation of the mappings W_k if the formant confidence was greater than a threshold (Table 2). The results show that as the threshold was increased, classification accuracy decreased. It appears that any benefit gained from only training on robust formant estimates was negated by the reduction in training data.

Including variance in model re-estimation. The final column of Table 1 shows the results of including an estimate of the matrix D_i in the re-estimation formulae (Eqs. 5 and 6), based on the expected value of E over vectors $\mathbf{x}(t)$ and $\mathbf{y}(t)$ (Eq. 3), from category k. The results show no significant advantage over ignoring D_i, and suggest that any benefit from including D_i is offset by inaccuracy in its estimation.

Visualization. Figure 2 (top) shows a spectrogram of an example speech utterance derived from its MFCCs, which indicates the richness of this parameterization. Below are its best approximations using a constant-trajectory SHMM (ID_0, center) and linear-trajectory MSHMM with the PFS12 intermediate layer (bottom). The extent of ID_0's success is perhaps surprising, in view of its piecewise stationarity. However, the MSHMM provides a much smoother rendering of the speech pattern, capturing many of its predominant characteristics (e.g., c. frames 20, 130 and 175). It is also better at transitions (e.g., frames 50–70, 110–120 and 215–230),

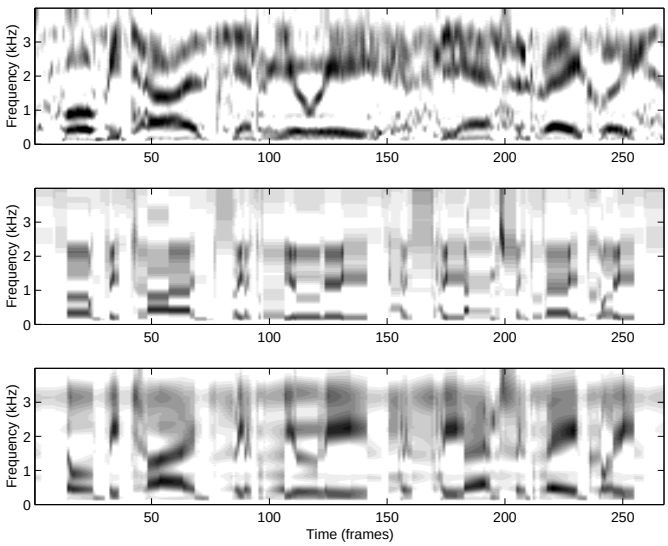

Fig. 2. Spectrograms of the utterance "Woe betide the interviewee if he answered vaguely" (`train/dr1/mcpm0/si1194.wav`) derived from (top) the original MFCCs, (center) constant-trajectory SHMMs, and (bottom) linear-trajectory MSHMM, using (c) 12 parallel-formant-synthesis control parameters and (F) 49 mappings.

although it is still poor in places (c. frames 100, 120 and 260), which may have been influenced by the lack of context-sensitivity in models of the consonants.

4.2. MOCHA experiments

Multi-level segmental HMM results. The results for each of the five jack-knifed sets, and the average across them, are presented in Table 3. Generally, the accuracy is higher than for the speaker-independent TIMIT tests, and again, with M = 10, the performance of the linear-trajectory models is superior to that of the constant-trajectory case. With six or more phone categories, the result is not significantly different from upper bound of the 13-dimensional linear-trajectory baseline.

5. CONCLUSIONS

A theoretical and experimental study of a simple class of multi-level segmental statistical models has been presented, in which speech dynamics are modeled as linear trajectories in an intermediate, articulatory-based representation, and mapped into acoustic space using a set of one or more linear transformations. It has been shown that with an appropriate combination of intermediate layer and number of transformations, performance near to the theoretical optimum can be achieved.

The significance of this result is that it provides a solid theoretical foundation for the development of richer classes of multi-level models, which include non-linear models of dynamics, alternative articulatory representations, sets of non-linear articulatory to acoustic mappings, and integrated optimization schemes that support unsupervised learning of the trajectory, intermediate representation and mapping parameters. The incorporation of a low dimensional, articulatory-based intermediate representation has many attractions. It provides a compact and interpretable framework for speaker adaptation and for meaningful characterization of the mechanisms which give rise to variability, such as in conversational speech, and for incorporation of articulatory constraints on the recognition process. It also has implications for model-based speech synthesis, and advances the goal of developing unified, trainable models which can support both recognition and synthesis.

In the future, we plan to investigate unsupervised training of both trajectory and articulatory-to-acoustic mapping parameters, to incorporate non-linear transformations, and in particular a radial basis function network, into the model, and to study the extent to which it is possible to increase the dimension of the acoustic parameterization, for example, by the addition of delta and acceleration parameters, while maintaining a compact intermediate layer.

6. REFERENCES

[1] Ostendorf, M., Digalakis, V., and Kimball, O.A., "From HMMs to segment models: a unified view of stochastic models for speech recognition", IEEE Trans. SAP, 4(5): 360–378, 1996.

[2] Jackson, P.J.B., Lo, B.-H., and Russell, M.J., "Data-driven, non-linear, formant-to-acoustic mapping for ASR", Electronics Letters (accepted May 2002).

[3] Richards, H.B., and Bridle, J.S., "The HDM: A segmental hidden dynamic model of coarticulation", Proc. IEEE-ICASSP'99, pp. 357–360, 1999.

[4] Deng, L., and Ma, J., "Spontaneous speech recognition using a statistical coarticulatory model for vocal-tract-resonance dynamics", J. Acoust. Soc. Am., 108(6): 3036–3048, 2000.

[5] Jackson, P.J.B., "Acoustic cues of voiced and voiceless plosives for determining place of articulation", Proc. CRAC workshop, pp. 19–22, Aalborg, Denmark, 2001.

[6] Holmes, W.J., and Russell, M.J., "Probabilistic-trajectory segmental HMMs", Computer Speech and Language, 13(1): 3–37, 1999.

[7] Holmes, J.N. , "Speech processing system using formant analysis", US patent 6292775, Sept. 2001.

[8] Wrench, A., "A new resource for production modelling in speech technology", Proc. Inst. Acoust., WISP'01, 23(3): 207–217, Stratford-upon-Avon, UK, 2001.