DEVELOPMENT OF ARTICULATORY-BASED MULTI-LEVEL SEGMENTAL HMMS FOR PHONETIC CLASSIFICATION IN ASR[†]

Martin J Russell[1], Philip J B Jackson[2] and Michael L P Wong[1]

[1]School of Engineering, University of Birmingham, UK. [m.j.russell@bham.ac.uk]
[2]Centre for Vision Speech and Signal Processing, University of Surrey, UK.

Abstract: *A simple multiple-level HMM is presented in which speech dynamics are modelled as linear trajectories in an intermediate, formant-based representation and the mapping between the intermediate and acoustic data is achieved using one or more linear transformations. An upper-bound on the performance of such a system is established. Experimental results on the TIMIT corpus demonstrate that, if the dimension of the intermediate space is sufficiently high or the number of articulatory-to-acoustic mappings is sufficiently large, then this upper-bound can be achieved.*

Key words: *Automatic speech recognition, Hidden Markov Models, segment models.*

1. INTRODUCTION

The goal of the research described in this paper is to develop efficient, complete and trainable acoustic models for speech processing that include an 'articulatory-based' representation, 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 coding and synthesis.

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 prone to error, and simple substitution with more-appropriate articulatory features is not viable. In the present work, states of the underlying Markov process 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.

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 [2]. Non-linear mappings, such as multi-layered perceptrons and radial-basis function networks, have been investigated [3], and many kinds of trajectory tried in the acoustic domain, e.g., [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 fixed linear-trajectory SHMM (of the type described in [6]) provides a theoretical upper bound on the performance of this type of MSHMM. Such a

[†]This work is funded by EPSRC grant M87146. See http://web.bham.ac.uk/p.jackson/balthasar/

model has been shown to outperform a conventional HMM [6]. Therefore, the goal of this paper is to determine whether this upper bound can be achieved by appropriate choice of articulatory representation and linear articulatory-to-acoustic mappings.

2. THEORY

2.1 Linear-trajectory segment models

The terminology follows [1] and [6]. Consider a fixed, linear-trajectory segmental HMM (FT-SHMM). Each state s_i of such a model is identified with a midpoint vector c_i and slope vector m_i, whose dimension N is that of the acoustic-feature space. A trajectory f_i of duration τ is defined by $f_i(t) = (t - \bar{t})m_i + c_i$, where $\bar{t} = (\tau + 1)/2$, and the probability of the sequence of acoustic vectors $y_1^\tau = \{y(1),..., y(\tau)\}$, given s_i, is

$$b_i(y_1^\tau) = \prod_{t=1}^{\tau} N(y(t); f_i(t), R_i), \tag{1}$$

where $N(y(t); f_i(t), R_i)$ is a multivariate Gaussian pdf with mean $f_i(t)$ and diagonal $N \times N$ covariance matrix R_i, evaluated at $y(t)$. The case $m_i = 0$ corresponds to a constant-trajectory SHMM [6].

2.2. Linear trajectories in the articulatory layer

Now consider a trajectory f_i in the M dimensional articulatory space. f_i is projected onto the acoustic layer by a mapping W, which is assumed to be linear. The midpoint $\boldsymbol{c}_i$ and slope $\boldsymbol{m}_i$ are M dimensional and equation (1) is replaced by:

$$b_i(y_1^\tau) = \prod_{t=1}^{\tau} N(y(t); W(f_i(t)), R_i). \tag{2}$$

2.3 Model parameter estimation

'Matched' articulatory-acoustic data are used to learn W. This is not necessary, but preserves the strict articulatory interpretation of the models in the intermediate layer. Given matched sequences a_1^τ and y_1^τ of articulatory and acoustic features, we use singular value decomposition to find a matrix W that minimizes the error,

$$E = \sum_{t=1}^{T} (Wa(t) - y(t))^T R_i^{-1} (Wa(t) - y(t)). \tag{3}$$

In general the speech is partitioned into K phone categories, each with a mapping W_k.

Let M be an S-state phone-level MSHMM, such that $a_{ij}=0$ if $j < i$. Then a state sequence x of length T can be written as $x = n_1 \otimes x_1, ..., n_R \otimes x_R$, where $R \le S$, $x_r=s_i$ for some i, and

$n_r \otimes x_r$ denotes n_r time frames in state x_r. A simple extension of the segmental Viterbi decoder ([6]) can be used to compute the state sequence $\hat{x}$ that maximizes:

$$P(y,x \mid M) = \pi(x_1) b_{x_1}\left(y_{t_1}^{t_2-1}\right) \prod_{r=2}^{R} a_{x_{r-1}x_r} b_{x_r}\left(y_{t_r}^{t_{(r+1)}-1}\right) \tag{4}$$

where the sequence x enters x_r at time t_r. Given $\hat{x}$, the maximum-likelihood estimates of the midpoint c_i and slope m_i for state s_i are

$$\hat{c}_i = \frac{1}{d_i} \sum_{t=t_i}^{t_{(i+1)}-1} \left(D_i W_k\right)^P D_i y(t) \text{ and } \hat{m}_i = \frac{\sum_{t=t_i}^{t_{(i+1)}-1} (t-\bar{t}) \left(D_i W_k\right)^P D_i y(t)}{\sum_{t=t_i}^{t_{(i+1)}-1} (t-\bar{t})^2} \tag{5}$$

respectively, where ^{P} denotes the pseudo-inverse, $D_i = R_i^{-\frac{1}{2}}$, $\bar{t} = \left(t_{(i+1)} + t_i - 1\right)/2$, $d_i = t_{(i+1)} - t_i + 1$ and k is the phone category for model M. If N=M and the rank of W_k is N, then $(D_i W_k)^P = W_k{}^P D_i{}^P$ and the D_i terms disappear from both equations. Interpreting equations (5), the optimal midpoint and slope parameters in the articulatory domain are those which give the give the best linear fit to the (pseudo) inverse-transformed observation vectors in the articulatory domain. If M=N and W_k is the identity mapping, then (5) coincide with the corresponding reestimation formulae for the slope and mid-point parameters in a FT-SHMM in [6].

2.4 Limitations of the linear model

It has been indicated elsewhere ([4]) that a linear transformation is inadequate for general articulatory-to-acoustic mapping. For example, consider the case of where speech is represented in the acoustic domain as the output of a set of D uniformly-spaced band-pass filters spanning frequencies up to 4kHz, and a single, hypothetical, 'formant' trajectory f, with unit amplitude, whose frequency increases linearly from 100Hz to 4kHz. The corresponding trajectory in acoustic space is a complex path over the surface of the D dimensional unit sphere, which passes through each of the axes in turn. This cannot be realised as the image of f under a linear mapping.

3. METHOD

3.1. Speech Data

All of the experiments use the TIMIT speech corpus. Speech from all male subjects in the TIMIT training and test sets was downsampled to 8kHz for compatibility with the formant analyser. Acoustic features (13 MFCCs including zeroth) were obtained using HTK (25ms window, 10ms fixed frame rate), while formant-based parameters for the intermediate layer were extracted using the Holmes formant analyser [7]. Three such parameterisations were considered: (a) first 3 formant frequencies (25Hz resolution); (b) first 3 formant frequencies plus 5 frequency-band energies; (c) the 12 control parameters from Holmes-Mattingly-

Shearme parallel formant synthesizer. A bias input (set equal to 1) was added to all of them to allow an offset to be learnt, for each acoustic feature. The data was partitioned into three sets: a *training set*, comprising speech from all male speakers in the TIMIT training set except for the first speaker in each dialect region, an *evaluation set*, comprising all of the speech from the first male speaker in each of the eight dialect regions, and a *test set* comprising speech from all male speakers in the TIMIT test set.

Acoustic models were built for each of the normal 49 TIMIT phones. Linear 'articulatory'-to-acoustic mappings were estimated using matched sequences of formant-based and acoustic data. Given these mappings, equations (5) were applied to re-estimate the MSHMM parameters using segmental Viterbi alignment. The maximum state duration was set to 15 frames in all experiments ($\tau_{max} = 15$).

3.2. Phone categories

Five different partitions of the phone set were considered, labelled A, C, D, E and F [1]. With one mapping per category, a series of mappings W_k was obtained for each categorization of the phones (number of mappings): A - all data (1); C - linguistic categories (6); D - as in Deng and Ma [5] (10); E - discrete articulatory regions [3] (10); F - individual phones (49) (these are the categories from [1], except that the two-class categorisation, B, is not included in the current experiments).

4. RESULTS

4.1. Language model scale factor

A phone-level probabilistic bigram language model was estimated using all of the TIMIT label files in the training set. Since acoustic and language model probabilities are not necessarily compatible, it is common practice to apply a *language model scale factor*. Typically this factor is a power, to which the language model probabilities are raised, or, equivalently, a multiplicative factor in the log probability domain. The optimal value of the language model scale factor was estimated empirically on the evaluation set (see figure 1). A value of 10 is close to optimal in all cases, and was used in all subsequent experiments.

4.2. Performance on the TIMIT male test set

The complete set of results is shown in table 1. The 'baseline' results for a FT-SHMM with zero and non-zero slope are 65.08% and 66.93% respectively. In the following, 'FT-SHMM' will refer to the FT-SHMM with non-zero slope. It has already been noted that since the image of a linear trajectory in articulatory space under a linear articulatory-to-acoustic mapping is linear, any MSHMM of the type considered in this paper is functionally equivalent to a linear trajectory FT-SHMM. Therefore, the performance achieved by this type of MSHMM can always be matched or exceeded by that of an appropriate FT-SHMM. In practice the algorithms used to train these models may only find local optima, and so the superior performance of any particular linear trajectory FT-SHMM cannot be guaranteed. However, table 1 shows that in these experiments the performance of the FT-SHMM is greater than that of the various MSHMMs in all cases.

As in [1], increasing the dimension of the intermediate representation, or the number of mappings, leads to improved results. The NIST implementation of the Matched Pair Sentence Segment (Word Error) Test [8] was used to assess the significance of differences between the performance of the FT-SHMM (66.93%) and that of each of the MSHMMs. For column (a), with 3 formant frequencies in the intermediate representation, the performance of all MSHMMs is significantly worse than the FT-SHMM. However, for column (b), where the intermediate representation also includes 5 band energies, the performance for 49 phone classes is statistically the same as that of the FT-SHMM. Finally, for an intermediate representation comprising 12 synthesiser control parameters, the performances for 1, 10(E) and 49 phone classes, and the FT-SHMM are statistically the same.

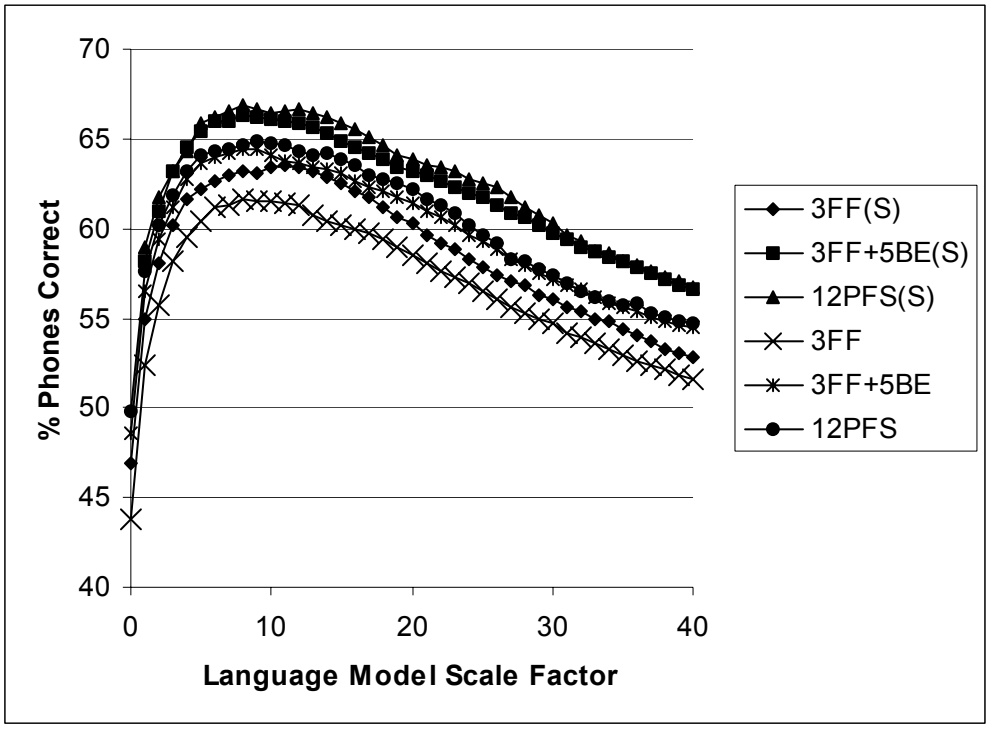

Figure 1: Effect of language model scale factor for MSHMMs with intermediate layer comprising 3 formant frequencies (3FF), 3 formant frequencies plus 5 band energies (3FF+BE) and 12 synthesiser control parameters (12PFS), where (S) indicates non-zero slope in the linear trajectories. Each graph is an average over the results for the five partitions of the phone set A(1), C(6), D(10), E(10) and F(49).

5. CONCLUSION

A simple multiple-level HMM has been presented in which speech dynamics are modelled as linear trajectories in an intermediate, formant-based representation and the mapping between the intermediate and acoustic data is achieved using one or more linear transformations. It is noted that such a system is functionally equivalent to an acoustic FT-SHMM with linear trajectories. Thus, by comparing the performances of a FT-SHMM and those of a range of MSHMMs, it is possible to measure the consequence of introducing an intermediate layer. Experimental results on the TIMIT corpus demonstrate that, if the dimension of the intermediate space is

sufficiently high or the number of articulatory-to-acoustic mappings is sufficiently large, then there is no significant difference between the performance obtained with a MSHMM and a FT-SHMM. Since linear transformations are inadequate for general formant-to-acoustic mapping, the promising results for these 'linear' MSHMMs suggests that future research into MSHMMs with non-linear articulatory-to-acoustic mappings is likely to be fruitful.

Mapping	Baseline	(a) F1-3	(b) F1-3 + BE5	(c) PFS12
ID (zero slope)	65.08			
ID (non-zero slope)	66.93			
A(1)		61.40	65.64	66.86*
C(6)		62.85	66.21	66.68
D(10)		62.57	66.43	66.19
E(10)		63.16	66.31	66.92*
F(49)		65.83	66.75*	66.92*

Table 1: Results on the test set (% phones correct). Column 2 refers to a standard FT-SHMM [6] with zero and non-zero slope. Columns 3, 4 and 5 refer to MSHMMs with intermediate representations comprising (a) 3 formant frequencies, (b) 3 formant frequencies plus 5 band energies, and (c) 12 synthesiser control parameters. Cases where there is no significant difference between the MSHMM and ID (non-zero slope) results are indicated by *.

REFERENCES

[1] P J B Jackson and M J Russell, Models of speech dynamics in a segmental-HMM recogniser using intermediate linear representations, *Proc. ICSLP 2002*, Denver, CO, 2002, 1253-1256.

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

[3] P.J.B. Jackson, B.H. Lo and M.J. Russell, Data-driven, non-linear, formant-to-acoustic mapping for ASR, *IEE Electronics Letters, Vol. 38, No. 13, 20th June 2002,* 667-669.

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

[5] L. Deng and J. Ma, Spontaneous speech recognition using a statistical coarticulatory model for vocal-tract-resonance dynamics', *J. Acoust. Soc. Am., 108(6): 2002*, 3036-3048.

[6] W.J. Holmes and M.J. Russell, Probabilistic-trajectory segmental HMMs, *Computer Speech and Language 1999*, 3-37.

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

[8] National Institute of Standards and Technology, Speech Group, Benchmark tests, http://www.nist.gov/speech/tests/sigtests/sigtests.htm.