

The Effect of an Intermediate Articulatory Layer on the Performance of a Segmental HMM

Martin J. Russell and Philip J. B. Jackson

School of Engineering, University of Birmingham, UK
Centre for Vision, Speech and Signal Processing, University of Surrey, UK

m.j.russell@bham.ac.uk

Abstract

We present a novel multi-level HMM in which an intermediate ‘articulatory’ representation is included between the state and surface-acoustic levels. A potential difficulty with such a model is that advantages gained by the introduction of an articulatory layer might be compromised by limitations due to an insufficiently rich articulatory representation, or by compromises made for mathematical or computational expediency. This paper describes a simple model in which speech dynamics are modelled as linear trajectories in a formant-based ‘articulatory’ layer, and the articulatory-to-acoustic mappings are linear. Phone classification results for TIMIT are presented for monophone and triphone systems with a phone-level syntax. The results demonstrate that provided the intermediate representation is sufficiently rich, or a sufficiently large number of phone-class-dependent articulatory-to-acoustic mapping are employed, classification performance is not compromised.

1. Introduction

This paper presents an empirical evaluation of a novel multi-level segmental hidden Markov model (MSHMM) in which the relationship between symbolic and acoustic representations of a speech signal is regulated by an intermediate ‘articulatory’ layer. In principle such a model has many advantages. For example, speech dynamics, which typically exhibit movement between frequency bands in the acoustic domain, can be modelled more effectively in an articulatory domain. Using such an approach it might also be possible to characterise the articulatory strategies that occur in fluent, conversational speech, or the physiological differences between an adult’s vocal tract and that of a child. An overview of segmental HMMs is presented in [2].

In our model, states of the underlying Markov process are associated with trajectories in an articulatory-based feature space (the intermediate layer). These trajectories are mapped into the acoustic feature space by an articulatory-to-acoustic mapping, where comparison is made with observations (figure 1). Of course, a potential problem with such a model is that any advantage gained by the introduction of an intermediate layer may be compromised by inadequacies of the articulatory representation, limitations of the articulatory-to-acoustic mapping, or theoretical compromises made for mathematical or computational tractability.

We consider a simple class of MSHMM whose trajectories in the articulatory-based representation are linear, and whose articulatory-to-acoustic mapping is realised as a set of one or

more linear mappings. We refer to such a model as a *linear-linear* MSHMM. Since the resulting trajectories in the acoustic-feature space are also linear, the performance of an appropriate fixed linear-trajectory acoustic segmental HMM [5] provides a theoretical upper bound on the performance of this type of MSHMM. All of the intermediate representations we consider here are based on formant frequencies. The simplest intermediate representation consists of the first three formant frequencies, while the most complex comprises the twelve synthesiser control parameters from the Holmes-Mattingly-Shearme (HMS) formant synthesizer [3]. Although these representations are implicitly, rather than explicitly, ‘articulatory’, they will be referred to as ‘articulatory’ throughout this paper.

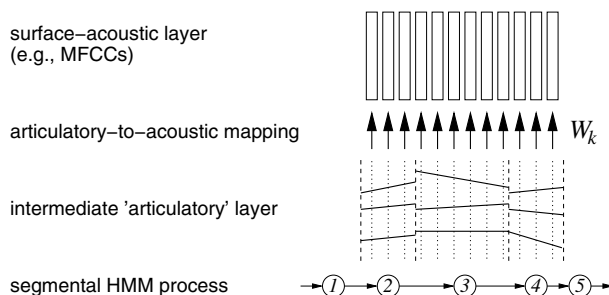


Figure 1: Illustration of a segmental model using linear trajectories in the intermediate space and mapping function W_k .

At this point it is worth noting that a linear-linear system is inadequate for speech pattern modelling [6]. Consider a case where speech is represented in the acoustic domain as the output of a set of A uniformly-spaced band-pass filters spanning frequencies up to 4 kHz, and a single, hypothetical, ‘formant’ trajectory f , with unit amplitude, whose frequency increases linearly from 100 Hz to 4 kHz. The corresponding trajectory in acoustic space is a complex path over the surface of the A dimensional unit sphere, which passes through each of the axes in turn. Such a trajectory cannot be realised as the image of f under a linear mapping.

This paper demonstrates that, even with the simple linear-linear system considered here, the upper bound on performance can be achieved by appropriate choice of articulatory representation and articulatory-to-acoustic mappings. It has been shown elsewhere that a fixed linear-trajectory SHMM can outperform a conventional HMM [5]. Hence the results presented in this paper give confidence that substantial improvements in performance relative to a conventional HMM can be achieved through the use of appropriate non-linear trajectories or non-linear articulatory-to-acoustic mappings.

The Balthasar project was funded by EPSRC grant M87146. See <http://web.bham.ac.uk/p.jackson/balthasar> for details.

2. MSHMM theory

2.1. Linear trajectories in an articulatory layer

The terminology follows [1]. Let \mathcal{M} be a fixed, linear-trajectory MSHMM. Each state s_i of \mathcal{M} is identified with midpoint and slope vectors c_i and m_i , respectively, whose dimension I is that of the intermediate articulatory space. A trajectory f_i of duration τ is defined by $f_i(t) = (t - \bar{t})m_i + c_i$, where $\bar{t} = \frac{\tau+1}{2}$. The probability of the sequence of A -dimensional acoustic vectors $y_1^\tau = y(1), \dots, y(\tau)$, given s_i , is

$$b_i(y_1^\tau) = \prod_{t=1}^{\tau} \mathcal{N}_{(W(f_i(t)), R_i)}(y_t), \quad (1)$$

where the $A \times I$ matrix W is a linear articulatory-to-acoustic mapping, R_i is an $A \times A$ (acoustic) covariance matrix, and $\mathcal{N}_{(\mu, R)}$ denotes a multivariate Gaussian density with mean μ and covariance matrix R .

2.2. Model parameter estimation

Now suppose that phones are partitioned into K categories. For each phone category k a separate articulatory-to-acoustic transformation W_k is learnt using ‘matched’ articulatory and acoustic data corresponding to category k . Given such matched sequences a_1^τ and y_1^τ of articulatory and acoustic features, we use singular value decomposition to find a matrix W_k that minimizes the error:

$$E = \sum_{t=1}^{\tau} (W_k a(t) - y(t))^T R^{-1} (W_k a(t) - y(t)). \quad (2)$$

If \mathcal{M} is an \mathcal{S} -state phone-level MSHMM, such that $a_{ij} = 0$ if $j \leq i$, then a state sequence x of length T can be written as $x = d_1 \otimes x_1, \dots, d_J \otimes x_J$, where $J \leq \mathcal{S}$, $x_r = s_i$ for some i , and $d_j \otimes x_j$ denotes d_j time frames in state x_j . Viterbi decoding can be used to compute the state sequence \hat{x} that maximizes:

$$p(y, x | \mathcal{M}) = \pi(x_1) b_{x_1}(y_{t_1}^{t_2-1}) \prod_{j=2}^J a_{x_{j-1}x_j} b_{x_j}(y_{t_j}^{t_{j+1}-1}), \quad (3)$$

where the sequence x enters state x_j at time t_j . Given \hat{x} , the maximum likelihood estimates \hat{m}_i and \hat{c}_i of the slope and midpoint for state s_i are:

$$\hat{m}_i = \frac{\sum_{t=t_i}^{t_{i+1}-1} (t - \bar{t})(D_i W_k)^\dagger D_i y(t)}{\sum_{t=t_i}^{t_{i+1}-1} (t - \bar{t})^2} \quad (4)$$

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

where X^\dagger denotes the pseudo-inverse of a matrix X , $D_i = R_i^{-\frac{1}{2}}$, $\bar{t} = \frac{t_{i+1} + t_i}{2}$, $d_i = t_{i+1} - t_i + 1$ and k is the phone category for model \mathcal{M} . If $A = I$ (so the dimensions of the articulatory and acoustic vectors are the same) and W_k is invertible then the D_i terms disappear from both equations. Interpreting equations 4 and 5, the optimal midpoint and slope parameters in the articulatory domain are those which give the best linear fit to the (pseudo) inverse-transformed observation vectors in the articulatory domain. If $A = I$, the number of phone categories is 1 (i.e., $K = 1$) and W_1 is the identity mapping, then equations 4 and 5 reduce to the corresponding reestimation formulae for the slope and mid-point parameters in a FT-SHMM in [5].

3. Experimental method

3.1. Speech data

All of the experiments reported here use the TIMIT speech corpus. Data from all of the male subjects in the TIMIT training and test sets was downsampled to 8 kHz for compatibility with the formant analyser. Acoustic features (13 MFCCs including zeroth) were obtained using HTK (25 ms window, 10 ms 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 (25 Hz resolution); (b) first 3 formant frequencies plus 5 frequency-band energies; (c) the 12 control parameters from HMS 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 were 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.

3.2. Phone categories

Five different partitions of the phone set were considered, corresponding to categories A, C, D, E and F from [1]. With K categories and one mapping per category, a series of mappings W_k , $k = 1, \dots, K$ was obtained for each categorisation of the phones: A - all data (1 mapping); C - linguistic categories (6 mappings); D - as in Deng and Ma [4] (10 mappings); E - discrete articulatory regions [8] (10 mappings); F - individual phones (49 mappings) [1]. Note that B, the two-class categorisation from [1], was not included in the current experiments.

3.3. Monophone and triphone model sets

The parameters of a set of 49 conventional, monophone acoustic HMMs were estimated for the TIMIT phone set using the tools in HTK [10]. In all cases it was assumed that the state covariance matrices were diagonal. For each conventional HMM \mathcal{M} , representing a phone in class k , a MSHMM \mathcal{M}' was created as follows:

- The mid-point vector c_i for the i^{th} state of \mathcal{M}' in the articulatory domain was defined as $c_i = W_k^\dagger \mu_i$, where μ_i is the mean vector for the i^{th} state of \mathcal{M} .
- The slope vector m_i was set to zero.
- The variance vector r_i for the i^{th} state of \mathcal{M}' in the acoustic domain was set equal to the variance vector for the i^{th} state of \mathcal{M} .
- The transform W_k and its pseudo-inverse W_k^\dagger were appended to the model.

Given these initial models, Viterbi alignment and equations 4 and 5 were used to re-estimate the MSHMM state parameters. The maximum state duration was set to 15 frames ($\tau_{\max} = 15$). These monophone MSHMMs were used to seed a set of triphone MSHMMs, which were again (locally) optimised using equations 4 and 5. The triphone set was defined by a simple ‘backoff’ scheme driven by a minimum-count parameter n_{min} , which ensured sufficient training examples to estimate the triphone parameters. A triphone MSHMM was created if at least n_{min} examples of the relevant phone-in-context occurred in the training set, otherwise the corresponding left-context biphone

was used instead. If the number of examples of this biphone context in the training set was less than n_{min} , then the monophone MSHMM was used instead. Small values of n_{min} will lead to large numbers of models and more accurate modelling of contextual effects. However, if n_{min} is too small there will be insufficient training data for robust training.

3.4. Language model

A phone-level probabilistic bigram language model was estimated using the TIMIT label files for data 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 λ . Typically the language model scale factor is multiplicative in the log probability domain.

4. Results

4.1. Monophone results

Table 1 shows phone classification results on the male TIMIT test set obtained using the various monophone systems. In all cases the language model scale factor was set to 10 ($\lambda = 10$), as this was demonstrated to be optimal on the evaluation set for almost all of the monophone systems [9].

The ‘baseline’ results for a monophone FT-SHMM with zero and non-zero slopes (ID0(m) and ID1(m) in table 1) are 65.08% and 66.93% respectively. Because the image of a linear trajectory 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, of course, 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 was 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, led to improved results.

The NIST implementation of the Matched Pair Sentence Segment (Word Error) Test [11] 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 was significantly worse than the FT-SHMM ID1(m). However, for column (b), where the intermediate representation also included 5 band energies, the performance for 49 phone classes was 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 were not statistically different.

Clearly, good classification performance depends on the ability of the linear mapping to recover an accurate acoustic representation from the articulatory data. The monophone results show that this can be achieved by using a sufficiently rich articulatory representation, such as the set of HMS synthesiser control parameters, in which case the number of phone categories appears to be unimportant, or by using a poorer articulatory representation and capturing phone-class dependent properties using a larger number of phone-class dependent articulatory-to-acoustic mappings.

Table 1: *Phone classification results (%) using monophone MSHMMs. ID0(m) and ID1(m) are acoustic (i.e., no intermediate layer) fixed trajectory monophone SHMMs with zero and non-zero slopes, respectively.*

| Map | Base | (a) | (b) | (c) |
|--------|-------|----------|-------|-----|
| | F1-3 | F1-3+BE5 | PFS12 | |
| ID0(m) | 65.08 | | | |
| ID1(m) | 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 | |

4.2. Triphone system parameters

In order to construct a set of triphones, it was necessary to determine an appropriate value for the ‘back-off’ parameter n_{min} . Figure 2 shows phone classification accuracy on the evaluation set as a function of this parameter. Surprisingly, classification accuracy was maintained, or even improved, even for very small values of n_{min} . However, these small values result in a large number of models and hence in excessive computational load. However, the Matched Pair Sentence Segment Test indicates that there was no significant difference between the performance obtained with n_{min} set to 30 and that obtained with smaller values. Hence n_{min} was set to 30, giving a set of exactly 1 400 triphones.

A further set of phone classification experiments was conducted on the evaluation set to determine an appropriate value for the language model scale factor, λ , for a triphone phone classification system, and a value of $\lambda = 6$ was found to be optimal.

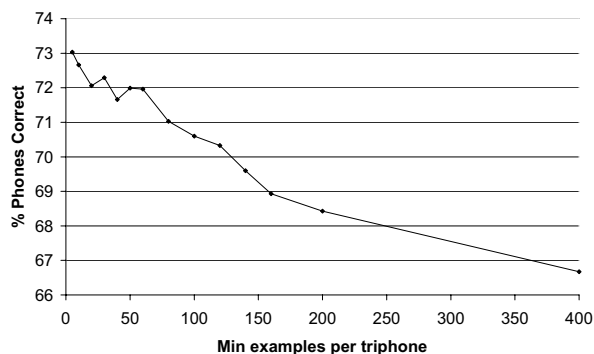


Figure 2: Phone classification accuracy on the TIMIT male evaluation set for different values of n_{min} .

4.3. Triphone results

Table 2 shows phone classification accuracy achieved when various triphone MSHMM systems were evaluated on the male portion of the TIMIT evaluation set. The ‘baseline’ result for a triphone FT-SHMM with non-zero slopes (ID1(t) in table 1) is 72.43%. As in the monophone case, this was a theoretical upper bound for the performances of all of the ‘linear-linear’

triphone MSHMM systems. The third column of table 2 corresponds to an intermediate representation with just 3 formant frequencies. As with the monophone systems, with the exception of phone categorisation D, classification accuracy increased with the number of categories.

The final column of table 2 shows phone classification results for triphone MSHMMs in which the intermediate representation consisted of the 12 PFS control parameters from the HMS parallel formant synthesizer. In this case the Matched Pair Sentence Segment Test indicated that the differences between the baseline FT-SHMM performance and the performance of any of the MSHMM systems were not statistically significant. However, the performances achieved with both the A(1) and F(49) phone categories are significantly better than the performance for the C(6) case.

Finally, table 3 shows TIMIT phone classification accuracy on the TIMIT male test set using the 3FF and 12PFS intermediate representations, (a) and (c) respectively, and phone categories A, C, D, E and F.

Table 2: *Phone classification results (%) on the evaluation set using triphone MSHMMs. ID1(t) is an acoustic (i.e. no intermediate layer) fixed trajectory triphone SHMM with non-zero slope.*

| Map | Base | (a) | (c) |
|--------|-------|-------|-------|
| | | F1-3 | PFS12 |
| ID1(t) | 72.43 | | |
| A(1) | | 67.23 | 72.03 |
| C(6) | | 68.00 | 71.36 |
| D(10) | | 67.53 | 71.99 |
| E(10) | | 68.33 | 71.96 |
| F(49) | | 70.15 | 72.49 |

Table 3: *Phone classification results (%) on the TIMIT male test set using triphone MSHMMs.*

| Map | (a) | (c) |
|-------|-------|-------|
| | F1-3 | PFS12 |
| A(1) | 67.21 | 72.78 |
| C(6) | 67.46 | 72.31 |
| D(10) | 67.37 | 72.01 |
| E(10) | 68.05 | 72.75 |
| F(49) | 70.32 | 72.70 |

5. Conclusions

This paper has described a novel multi-level segmental HMM (MSHMM) which incorporates an intermediate ‘articulatory’ layer between the state and surface-acoustic layers. In the case where the articulatory-to-acoustic mapping is linear, the performance of such a MSHMM is bounded above by that of an appropriate acoustic fixed-trajectory segmental HMM, and so the effect of introducing an intermediate layer on performance can be measured. Experimental results on the male subjects in the TIMIT test set have been presented for simple monophone and triphone MSHMM systems of this type. Even for this simple case, it has been shown that (provided the system uses either a sufficiently large number of phone-class-dependent linear mappings, or a sufficiently rich articulatory representation) optimal performance can be achieved.

It has been pointed out that linear mappings are inadequate for articulatory-to-acoustic mapping. Hence the results presented here provide compelling motivation for the development of MSHMMs with non-linear articulatory-to-acoustic mappings. Research in this area is currently being pursued, in which the articulatory-to-acoustic mapping is achieved using both generic and customised artificial neural networks. In these cases, the mathematics of model parameter estimation is more complex, however progress is being made.

6. References

- [1] Jackson, P. J. B., and Russell, M. J., “Models of speech dynamics in a segmental-HMM recognizer using intermediate linear representations”, Proc. ICSLP 2002, Denver, CO, 1253-1256, 2002.
- [2] 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.
- [3] Holmes, J. N., Mattingly, I. G., and Shearme, J. N., “Speech synthesis by rule”, Language and Speech, 7, pp 127-143, 1964.
- [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, 2002.
- [5] Holmes, W. J. and Russell, M. J., “Probabilistic-trajectory segmental HMMs”, Computer Speech and Language 13(1), 3-37, 1999.
- [6] Richards, H. B. and Bridle, J. S., “The HDM: A segmental hidden dynamic model of coarticulation”, Proc. IEEE-ICASSP’99, 357-360, 1999.
- [7] Holmes, J. N., “Speech processing system using formant analysis”, US patent 6292775, Sept. 2001.
- [8] Jackson, P. J. B., Lo, B. H., and Russell, M. J., “Data-driven, non-linear, formant-to-acoustic mapping for ASR”, IEE Electronics Letters, 38(13), 667-669, 2002.
- [9] Russell, M. J., Jackson, P. J. B., and Wong, L. P., “Development of articulatory-based multi-level segmental HMMs for phonetic classification in ASR”, to appear in Proc. EC-VIP-MC 2003, 4th EURASIP Conference on Video/Image Processing and Multimedia Communications, Zagreb, Croatia, 2003.
- [10] Young, S., Odell, J., Ollason, D., Valtchev, V. and Woodland, P., “The HTK Book”, Cambridge University, 1997.
- [11] National Institute of Standards and Technology (Speech Group), “Benchmark tests”, Gaithersburg, MD, <http://www.nist.gov/speech/tests/sigtests/sigtests.htm>, 2000.