Speech recognition using non-linear trajectories in a formant-based articulatory layer of a multiple-level segmental HMM

Hongwei Hu, Martin J. Russell

Department of Electronic, Electrical and Computer Engineering, University of Birmingham, Birmingham B15 2TT, UK

hwh400@bham.ac.uk, m.j.russell@bham.ac.uk

Abstract

This paper describes how non-linear formant trajectories, based on ‘trajectory HMM’ proposed by Tokuda et al., can be exploited within the framework of multiple-level segmental HMMs. In the resultant model, named a non-linear/linear multiple-level segmental HMM, speech dynamics are modeled as non-linear smooth trajectories in the formant-based intermediate layer. These formant trajectories are mapped into the acoustic layer using a set of one or more linear mappings. The N-best rescoring paradigm is employed to evaluate the performance of the non-linear formant trajectories. The rescoring results on TIMIT corpus show that the introduction of non-linear formant trajectories results in improvement on recognition phone accuracy compared with linear trajectories. 

Index Terms: speech recognition, non-linear formant trajectories, segmental HMMs

1. Introduction

The purpose of this paper is to determine whether the use of smooth, non-linear formant trajectories in a multiple-level segmental HMM (MSHMM) can result in improved speech recognition performance compared with a MSHMM using linear formant trajectories. In a multiple-level linear/linear segmental HMM previously presented in [1], the relationship between the underlying symbolic and surface acoustic (e.g. MFCCs) representations of a speech signal is regulated by an intermediate ‘articulatory’ layer. Figure 1 (a) shows the structure of a linear/linear MSHMM. States of the underlying Markov process are associated with piecewise linear trajectories in the intermediate layer, which are mapped into the acoustic layer using a linear articulatory-to-acoustic mapping. The results of phonetic classification experiments on TIMIT show that, even with this simple linear/linear MSHMM system, speech recognition performance can achieve the upper bound of an appropriate fixed linear-trajectory acoustic segmental HMM (FT-SHMM), which, in turn, can outperform a conventional HMM [2]. It was hoped that further improvements in performance relative to a conventional HMM could be achieved by using appropriate non-linear models of dynamics, alternative articulatory representations or non-linear articulatory-to-acoustic mappings.

One of the limitations of the linear/linear MSHMM is the use of linear trajectories. A linear trajectory is simply characterized by a mid-point and a slope vector, which are of the same dimension of the intermediate layer. No continuity constraints are applied across the segment boundaries. Although a piecewise linear model provides an adequate ‘passive’ approximation to the formant trajectories, it does not capture the active dynamics of the articulatory system. Many alternative intermediate-layer models of dynamics have been proposed, e.g., [3, 4, 5, 6, 7, 8]. In addition, it has been noted that a linear articulatory-to-acoustic mapping is inadequate in general [3].

In this paper, non-linear formant trajectories are generated based on Tokuda’s ‘trajectory HMM’ method [9]. For consistency with the linear/linear MSHMM described in [1], the articulatory-to-acoustic mappings are also linear, but the mapping parameters are re-estimated based on the non-linear trajectories data. The resultant model is referred to as a non-linear/linear MSHMM, and is shown in Figure 1(b).

Five ‘formant-to-acoustic’ mapping schemes, based on different phone categories [1, 10], are considered in order to investigate their effects on the system performance. Phonetic recognition experiments on TIMIT are performed to compare the performance of non-linear trajectories and linear trajectories. As an appropriate decoder for non-linear/linear MSHMMs is not yet available, an N-best list rescoring paradigm, where \( N = 1,000 \), is employed to evaluate the effectiveness of the non-linear formant trajectories.

The rest of this paper is organized as follows. Section 2 describes the formant-based intermediate layer. Section 3 briefly explains the speech parameter generation algorithm to produce non-linear trajectories. Speech recognition experiments and results are shown in Section 4 and 5. Conclusions and future work are presented in the final section.
2. Formant-based intermediate layer

The intermediate layer presented in [1] is based on formant frequencies, ranging from the simplest form which just consists of the first three formant frequencies to a 12 dimensional representation using the 12 parallel formant synthesiser (PFS) control parameters. Experimental results show that the performance improves either as the dimension of the intermediate representation or the number of mappings is increased [10]. This paper concentrates on the 12 PFS control parameter representation, referred to as 12PFS, which are produced by Holmes-Mattingley-Shearne (HMS) formant analyser [11]. The 12 PFS control parameters are listed in Table 1.

<table>
<thead>
<tr>
<th>No.</th>
<th>Parameter description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FN: Freq. of ‘low freq.’ formant (default 250Hz)</td>
</tr>
<tr>
<td>2</td>
<td>ALF: Amplitude of FN in dB</td>
</tr>
<tr>
<td>3</td>
<td>F1: Freq. of first formant (in 25 Hz steps)</td>
</tr>
<tr>
<td>4</td>
<td>A1: Amplitude of first formant in dB</td>
</tr>
<tr>
<td>5</td>
<td>F2: Freq. of second formant (in 50 Hz steps)</td>
</tr>
<tr>
<td>6</td>
<td>A2: Amplitude of second formant in dB</td>
</tr>
<tr>
<td>7</td>
<td>F3: Freq. of third formant (in 50 Hz steps)</td>
</tr>
<tr>
<td>8</td>
<td>A3: Amplitude of third formant in dB</td>
</tr>
<tr>
<td>9</td>
<td>AHF: Amp. in high freq. region in dB</td>
</tr>
<tr>
<td>10</td>
<td>V: Degree of voicing</td>
</tr>
<tr>
<td>11</td>
<td>F0: Fundamental freq. on logarithmic scale</td>
</tr>
<tr>
<td>12</td>
<td>MS: Glottal-pulse mark-space ratio</td>
</tr>
</tbody>
</table>

Table 1: The 12 parallel formant synthesiser (PFS) control parameter representation, 12PFS, in the intermediate layer of a MSHMM

It has to be confessed that this is a relatively simple ‘articulatory’ representation compared to those presented in [3, 4, 5, 6, 7, 8]. Also, strictly speaking, it is only implicitly articulatory by nature. However, a major motive of incorporating such an intermediate layer is to allow speech dynamics to be modelled simply and directly in an articulatory-related space, which is capable of supporting both recognition and synthesis. Indeed, it has been demonstrated that given a sequence of appropriate PFS control parameters, extremely high quality speech can be produced using a parallel formant synthesiser. This suggests that a MSHMM, whose intermediate layer is the 12PFS representation, can be used for trainable model-based speech synthesis. This motivates the concept of ‘unified’ model for speech recognition and synthesis proposed in [12].

3. Trajectory modelling

The non-linear formant trajectories in this paper are generated based on the ‘trajectory HMM’ method [9]. Most state-of-the-art automatic speech recognition (ASR) systems use static and dynamic features (e.g. delta and delta-delta coefficients) to accommodate temporal dynamics. Although the employment of dynamic features improves the performance of HMM-based speech recognizers, their use leads to inconsistencies in a conventional HMM, where it is assumed that both the static and dynamic parameters can be simultaneously constant and non-zero. The motivation for ‘trajectory HMM’ is to alleviate this inconsistency between static features and dynamic features of a conventional HMM. A trajectory, which is most consistent with the static and dynamic constraints is synthesised in the static feature vector space. In fact, the new trajectory is exactly the same as the speech parameter trajectory generated by using the speech parameter generation technique in [13]. A detailed description of the algorithm appears in [13], where it is referred to as case 1, but a brief explanation is given here for completeness.

Let \( O = \{ o_1, ..., o_T \} \) be a sequence of speech observations and \( S = \{ s_1, ..., s_T \} \) a fixed state sequence of an HMM \( M \). Assume that the speech vector \( o_t \) consists of both a static feature vector \( c_t \) and dynamic feature vectors \( \Delta c_t, \Delta^2 c_t \). The delta and delta-delta coefficients are computed using the following formulae, where \( \theta \) is set to 1 in this paper.

\[
\Delta c_t = \frac{c_{t+\theta} - c_{t-\theta}}{2\theta} \quad (1)
\]

\[
\Delta^2 c_t = \frac{\Delta c_{t+\theta} - \Delta c_{t-\theta}}{2\theta} \quad (2)
\]

Let \( W \) be the linear transform matrix which transforms the sequence of static parameter vectors \( C \) into an ‘augmented’ sequence of static plus dynamic vectors \( O \). Then the above formula can be written as

\[
O = WC \quad (3)
\]

For a given state sequence \( S \), the speech parameter generation problem is to determine the parameter sequence \( C \) which maximize \( P(O|S, M) \) with respect to \( C \) under the constraints (3). By setting

\[
\frac{\partial \log P(WC|S, M)}{\partial C} = 0 \quad (4)
\]

a set of equations are obtained. For detailed representation of the transform matrix \( W \) and equation (4), refer to [13]. The non-linear trajectories used in this paper are obtained by solving equation (4). This speech parameter generation technique forms the basis of HMM-based speech synthesis described in [13].

4. Experiment methods

4.1. Speech data

The male part of the TIMIT corpus is used for all experiments. The training data includes all male utterances from the TIMIT training set (3,252 utterances) and the test data includes utterances from male speakers in TIMIT core test set (128 utterances, excluding ‘sal’ and ‘sa2’ sentences). 13 MFCCs, including zeroth, are used as the acoustic features for MSHMMs training and evaluation. They are produced using HTK with 25 ms window, 10ms frame rate. 39 MFCCs, including \( \Delta \), \( \Delta^2 \) coefficients are also used in the experiments. The 12 PFS control parameters are produced by the HMS formant analyzer, converted to HTK format and augmented with \( \Delta \) and \( \Delta^2 \) coefficients, resulting in a representation of 36 dimensional features.

4.2. N-best list

An \( N \)-best list where \( N = 1,000 \) is generated using a set of standard decision-tree based triphone HMMs built using HTK, which consists of 1,000 most likely hypotheses for each test utterance. The triphone model set is built on 39 MFCCs (including \( \Delta \) and \( \Delta^2 \) coefficients). The baseline phone accuracy (1 best decoding) is 65.7%. The upper bound phone accuracy for the \( N \)-best list is 77.9%, which is based on the most accurate match from the 1,000 alternatives with the reference label files. Therefore, this is the theoretical maximum score the following \( N \)-best rescoring experiments can reach. The optimal state sequence for each \( N \)-best list hypothesis can be produced using forced alignment with HTK.
4.3. Model sets

Five linear/linear monophone MSHMM model sets are built on the TIMIT corpus using 13 MFCCs (static features only), based on different articulatory-to-acoustic mapping schemes. The MSHMM model parameters are optimized using an estimation-maximization (EM) scheme based on segmental Viterbi decoding [1], implemented as part of the ‘SEGVis’ toolkit.

A | all data
B | vowels, {hh,lr,wy}, nasals, {dh,fs,sh,th,v,z,zh} {ch,jh} {b,cl,d,vcl,cl,epi,g,cl,pl}
C | vowels, {epi,q,sil}, {hh,lr,wy}, nasals, {ch,s} {sh,fl,th}, {dh,v,z}, {j,j,z}, {cl,k,p,t}, {b,d,vcl,dx},
D | vowels, {epi,q,sil}, {dx,cl,lr,wy}, nasals, {vcl}, {cl} {b,d,g,jh}, {ch,k,p,q,t}, {dh,v,z,zh}, {fl,hh,s,sh,th}
E | 49 individual phones

Table 2: Definitions of the phone categories: B. linguistic categories; C. as in [6]; D. discrete articulatory regions. ‘nasals’ and ‘vowels’ denote the sets {en,em,ng}, and {aa,ae,ah,ao,aw,ax,ay,eh,el,er,ey,ih,ix,oy,ow,uh,uw} respectively.

The formant-to-acoustic mapping schemes considered in this paper include all five mapping schemes described in [1], ranging from a single ‘phone-independent’ mapping to 49 ‘phone-dependent’ mappings. The five mapping schemes are referred to as 1A, 6B, 10C, 10D and 49E respectively (number of mappings followed by phone partition), as shown in Table 2. The mappings are estimated by minimizing the error between matched sequences of formant and acoustic data. For example, assume that \( f = \{ f_1, ..., f_T \} \) and \( y = \{ y_1, ..., y_T \} \) are formant and acoustic sequences corresponding to a particular phone class, then the mapping \( W \) is calculated by minimizing the error \( E \) as follows:

\[
E = \sum_{t=1}^{T} \| W f_t - y_t \|^2
\]

In practice, the mappings are pre-computed using the TIMIT corpus before estimation of the MSHMM model parameters. A bias, which is set to ‘1’, is added to 12 PFS control parameters to accommodate offsets.

4.4. Segment probability calculation

Suppose that the unknown utterance corresponds to a sequence of acoustic feature vectors \( y = y_1, ..., y_T \). For a given entry in the N-best list, this utterance corresponds to a state sequence \( s = s_1, ..., s_T \) where \( s_1 + s_2 + ... + s_T = T \). A trajectory of length \( T \) is defined in the intermediate space as \( f = f_1, f_2, ..., f_T \), where \( f_1 \) is a trajectory ‘fragment’ of length \( T_1 \). In the case of a linear/linear MSHMM each \( f_n \) is a linear trajectory, defined by a slope and mid-point value for state \( s_n \), as in [1]. For a non-linear/linear MSHMM \( f \) is a single, continuous trajectory obtained using the ‘trajectory HMM’ method, and \( f_n \) is simply the section of \( f \) which corresponds to state \( s_n \). The probability of \( y \) is then given by:

\[
p(y) = \sum_{n=1}^{N} a_n(\tau_n) \prod_{i=1}^{T_n} \left[ N(y_{\phi_n+i}; W_n(f_n(y_{\phi_n+i})), \sigma_n) \right]
\]

where \( \phi_n = \sum_{i=1}^{n-1} \tau_i \), \( N(y; \mu, \sigma) \) denotes a multivariate Gaussian probability density function (PDF), with mean \( \mu \) and diagonal covariance \( \sigma \) and \( \sigma_n \), and \( d_n \) are the (acoustic) variance and duration PDF associated with state \( n \). The state duration PDF was uniform with the maximum state duration set to 15 frames \( (\tau_{max} = 15) \).

4.5. Rescoring with linear/linear MSHMMs

For comparison purpose, the baseline experiment is to rescoring the N-best list using linear/linear MSHMMs. Rescoring with linear/linear MSHMMs is quite straightforward. For a given state sequence, equation (6) is used to calculate the probability of a sequence of acoustic features. Each hypothesis in the N-best list is rescoring and the one with the highest probability is recorded for evaluation.

It is worth noting here that this paper considers two types of state sequence \( s_1, ..., s_N \). One is generated using conventional HMMS, with HTK. Phone boundaries are specified as a consequence of the forced alignment. This is referred to as a ‘constrained’ state sequence. In this case the MSHMM is forced to use the same state sequence. The other, ‘unconstrained’ state sequence is produced, ignoring the conventional HMM boundary timing, with ‘SEGVis’. Both are used in turn to rescoring the N-best list with linear/linear MSHMM. However, only the constrained state sequence is used in the non-linear rescoring experiment.

4.6. Rescoring with non-linear/linear MSHMMs

As the non-linear trajectories are realized in the articulatory-based space, a set of conventional monophone HMMS (single Gaussian) are built on 36 dimensional PFS control parameters (12 PFS control parameters plus \( \Delta \) and \( \Delta^2 \) coefficients) using HTK. To rescore with non-linear/linear MSHMMs, non-linear trajectories are generated for each N-best list hypothesis. The state sequence is a constrained state sequence, which is the same as that used in the constrained linear/linear MSHMMs rescoring. Given this state sequence, the non-linear trajectory can be directly generated using the technique described in Section 3.

5. Experiment Results

5.1. Linear/linear MSHMMs rescoring results

Tables 3 and 4 show the rescoring results for linear/linear MSHMMs, where the state sequence used in Table 3 is constrained and the unconstrained state sequence is used in Table 4. The third rows of the tables show the rescoring results after the (correct) reference transcriptions are added to the N-best lists.

<table>
<thead>
<tr>
<th>Mapping</th>
<th>1A</th>
<th>6B</th>
<th>10C</th>
<th>10D</th>
<th>49E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone accuracy</td>
<td>65.7</td>
<td>65.0</td>
<td>66.0</td>
<td>66.3</td>
<td>65.4</td>
</tr>
</tbody>
</table>

Table 3: Phone accuracy of rescoring with linear/linear MSHMMs using constrained state sequence on different mapping schemes

Table 3 shows that the MSHMM based on mapping ‘10D’ gives the highest phone accuracy, either without or with the reference transcriptions. However, for the unconstrained system Table 4 shows the highest phone accuracy is reached using ‘49E’ scheme. In case reference transcriptions are not included,
the unconstrained rescoring results (Table 4) are generally better than the corresponding constrained results. Furthermore, the unconstrained rescoring results in Table 4 also show the trend that phone accuracy increases as the number of articulatory-to-acoustic mappings increases.

### 5.2. Non-linear/linear MSHMMs rescoring results

<table>
<thead>
<tr>
<th>Mapping</th>
<th>1A</th>
<th>6B</th>
<th>10C</th>
<th>10D</th>
<th>49E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone accuracy</td>
<td>65.3</td>
<td>65.8</td>
<td>66.1</td>
<td>66.4</td>
<td>66.4</td>
</tr>
<tr>
<td>Phone acc. (N+1)</td>
<td>67.8</td>
<td>68.4</td>
<td>69.7</td>
<td>68.6</td>
<td>70.4</td>
</tr>
</tbody>
</table>

Table 5: Phone accuracy of rescoring with non-linear/linear MSHMMs using constrained state sequence on different mapping schemes. Mapping reest. means MSHMM mappings are re-estimated.

Table 5 shows the results of rescoring the N-best list using non-linear/linear MSHMMs. MSHMM(old) means the rescoring is based on the original MSHMM models, while ‘mapping reest.’ indicates that an updated MSHMM model set, whose articulatory-to-acoustic mapping parameters are re-estimated relative to the non-linear trajectories, is used for rescoring. The mapping re-estimation is also based on equation 5. The original mappings are trained based on the matched sequence of 12 PFS control parameters and 13 MFCCs in the TIMIT training set. However, instead of using the original 12 PFS control parameters, the mappings are re-trained based on a sequence of smooth non-linear trajectories data derived from using the same algorithm discussed in Section 3. For each utterance in the TIMIT training set, non-linear trajectories are generated in the same way with a different set of control parameters derived from the same algorithm.

6. Conclusions and future work

This paper has shown that the use of non-linear formant trajectories achieves a reduction in phone error rate on TIMIT, although the improvement is modest. Hence it can be argued that the use of non-linear formant trajectories improves the recognition performance, compared with piecewise linear trajectories. This paper also compares the performance of different articulatory-to-acoustic mapping schemes and the constrained and unconstrained state sequences in the rescoring experiments.

The model sets used in this paper comprise monophone MSHMMs. Of course, the results of a triphone experiment would be more interesting and this experiment is ongoing. It is hoped that rescoring the same N-best list using triphone MSHMMs would give a higher score of phone accuracy. However, the monophone results presented in this paper show that the MSHMM performance can be improved by using a richer class of non-linear smooth trajectories and more sophisticated articulatory-to-acoustic mappings.

In the future, a short term goal is to evaluate the non-linear formant trajectories using context-dependent triphone MSHMMs. The N-best rescoring paradigm will still be used at this stage. A longer-term goal is to develop a Viterbi type decoder for the training and evaluation of non-linear/linear MSHMMs.

7. References