Trajectory behaviour at different phonemic context sizes

Jaco Badenhorst
Multilingual Speech Technologies
North-West University,
Vanderbijlpark 1900, South Africa
[2]Human Language Technology
Competency Area,
CSIR Meraka Institute
Email: jbadenhorst@csir.co.za

Marelie H. Davel
Multilingual Speech Technologies
North-West University,
Vanderbijlpark 1900, South Africa
Email: marelie.davel@gmail.com

Etienne Barnard
Multilingual Speech Technologies
North-West University,
Vanderbijlpark 1900, South Africa
Email: etienne.barnard@gmail.com

***Abstract*—We propose a piecewise-linear model for the temporal trajectories of Mel Frequency Cepstral Coefficients during phone transitions. As with conventional Hidden Markov Models, the parameters of the model can be estimated for different phonemic context sizes, but our model allows for an intuitive understanding of the impact of context size. We find that the most detailed models, predictably, match the coefficient tracks best – but when data scarcity forces us to use less detailed models, different styles of context modelling (clustered triphones versus biphones) have complementary behaviours. We discuss how this complementarity may be useful for data-efficient ASR.**

I. Introduction

Hidden Markov Model (HMM)-based speech recognition systems are notoriously dependent on the availability of large amounts of training data. This data is also required to be phonetically rich: it is not sufficient to include a large number of samples of each monophone; these monophones must appear in the required contexts.

Modelling larger acoustic contexts (increasing the number of phonemes considered to the left and right of the unit being modelled) exponentially increases the number of acoustic models required. And as the number of models increases, so the need for additional data is increased. At a context size that is large enough to only combine models that are indeed similar, data typically becomes too sparse for accurate model estimation. To deal with this data sparsity, typical HMM systems employ tree-based clustering. States that are acoustically similar are grouped together, ensuring that adequate model estimation can be accomplished.

The success of triphones is at least partly a consequence of their flexibility near phone transitions. Given the physical constraints of the human vocal tract, a transition of one phonetic unit to the next is bound to be coupled by co-articulation. We are therefore interested in understanding whether some triphones display similar acoustic changes during phone transitions, and how well this behaviour can be approximated by smaller, less data-hungry units (such as biphones).

Since the co-articulation effect varies over time, we are particularly interested in understanding whether temporal information near phone transitions show systematic effects within different groupings of triphones. If this is the case, such information can be used to better predict the behaviour of rare or unseen triphones, based on the behaviour of similar triphones or even units with less context (biphones or monophones).

In this work, we present a model that can be used to isolate the key elements of the acoustic changes that occur in each phone-to-phone transition. We first show that trajectory behaviour in general can be modelled this way (that the different transitions in a set of acoustic data can be described with the new model). We then analyse transition behaviour by grouping the transitions in different ways and evaluating the accuracy with which these constrained models are still able to represent our data. We also discuss how the result of this analysis points to alternative methods that can be considered for constructing multi-unit models.

This paper is structured as follows: We discuss some related research in section II. Section III describes the specific techniques we apply to model and analyse temporal characteristics of phone transitions. We then describe our experimental set-up in section IV and present the results in section V. This is followed by a summary of our main observations in section VI.

II. Background

The importance of contextual modelling has long been understood [1]. In particular, tying triphones at the state level using a phonetic tree is considered an effective modelling approach, and is standard practise when building high-accuracy speech recognition systems [2], [3]. However, when two acoustic units are clustered together (to compensate for data sparsity), the scores returned by the acoustic model will always be the same, even when these units may be quite dissimilar acoustically. In response to this, Chang and Glass [4] proposed a back-off discriminative acoustic modelling method that incorporates broad phonetic classes. Their model requires specific acoustic-phonetic knowledge to subdivide the classification problem into sub-problems, and augments the overall acoustic scores with that of the sub-problems. A fully automated state-based Eigentriphones modelling approach is shown to be just as successful [5]. This procedure attempts to retain acoustic discrimination using the careful adaptation of HMM parameters.

Apart from the size and grouping of contexts, another key element of contextual modelling relates to the inclusion of temporal information in the model. When building HMM-based systems, first and second derivatives of the underlying features (such as MFCCs) are typically added to the set of features being modelled [3]. This results in a simple but effective technique for the modelling of temporal information.

A more explicit modelling of temporal effects may be required for accurate representation. Evidence from the speech production process suggests the existence of underlying articulatory trajectories in speech data [6]. As a result, much research in spoken language technology intends to incorporate structures of human speech into current statistical speech recognition systems [7].

Attempts at explicit modelling of temporal trajectories have achieved limited success [8]. Generally, these approaches attempt to overcome specific limitations of the HMM modelling paradigm (especially the state-based independence assumption), by either incorporating explicit trajectories within the HMM framework [9] or by defining longer-term variable-length segmental models [10].

While there is an extensive field of literature related to improving speech recognition accuracies for well-resourced languages, the implication of temporal modelling when working with systems trained on limited amounts of data, is not so clearly understood. [7] seem to obtain rather promising results, as do [11], who show that frame-based feature trajectories are informative on the nature of transitions for specific phone classes.

III. Approach

At the heart of our approach to analysing trajectory behaviour is a linear trajectory model that captures temporal changes (at the cepstral level) for every phone transition. This model can be applied to different phone classes and at various contextual levels, and the differences between the modelled trajectories and the actual speech data measured. By constraining the model in different ways (grouping certain transitions as if they were similar) and evaluating the effect this has on the accuracy of the model, we gain an understanding of the acoustic changes that take place during phone transitions.

The trajectory tracking technique consists of the following main elements: (1) Preparation of the input features used to describe each transition, (2) linear trajectory estimation using a linear trajectory model, (3) calculation of reference values as required by this model, and (4) model evaluation and analysis.

A. Feature preparation

HMM-based ASR systems encode the speech-signal using frame-based feature vectors such as Mel-Frequency Cepstral Coefficients (MFCCs) or Linear Predictive Coding (LPC) coefficients. We utilise MFCCs in the current analysis but other frame-based features would also be applicable within the general framework described here.

As a first step towards trajectory parameter estimation, phone transition boundaries are obtained using an ASR system in forced alignment mode. Guided by the estimated phone boundaries, we define specific phone transitions by segmenting all of the phone examples at their centres (which are expected to be the most stationary part of each phone) effectively yielding diphone units. These phone transition units can then be described by tying together their parameters at the mono-, bi- or triphone level. This results in a set of labelled transitions for each unit-to-unit pair.

B. Linear trajectory estimation

The authors of [11] showed evidence for the different types of co-articulatory mechanisms at work for various phone transitions, modelled using MFCC features. Quite generally, however, plots of the 13 MFCCs for the frames of phone transitions seemed to suggest a definite change near the phone transition, and very little change near the centre of phones. Consequently, analysing cepstral trajectories for these transitions should be tractable using simple linear models.

In order to model this behaviour we use piece-wise linear approximation. Three line pieces are used to fit the cepstral values of a single MFCC coefficient stream, using least-squares optimisation. We restrict the start and end line segments to be constant values (linear with zero slope), and model the transition between these two values with a straight line of variable slope. Furthermore, we require the constant line segments (the start and end line pieces) to be associated with at least θ frames serving as trajectory anchor points, with $\theta = 1$ in the current work.

Estimation of the centre line piece is not explicitly associated with any data points. Rather, we utilise the zero order anchor points and draw the first order line between the end and starting indexes of the two anchor points. We search for these indexes by optimising the squared error SE across all three line segments. This also yields a single error value for the specific approximation.

Finally, in order to compare the "goodness of approximation" for different options we calculate the square errors (SE_f) followed by the mean square error (MSE_{coef}):

$$SE_f = |t(x_f) - y_f|^2 \quad (1)$$

where $t(x_f)$ is the trajectory value at frame x_f and $|t(x_f) - y_f|^2$ is the squared residual.

$$MSE_{coef} = \frac{1}{F}\sum_{f=1}^{F} SE_f \quad (2)$$

Once optimised, this model then provides the following values:

ref_{start}	parameter value at initial stable point	
f_{start}	frame at start of the transition	
f_{end}	frame at end of the transition	
ref_{end}	parameter value of final stable point	(3)

As these are calculated for each coefficient individually, and can be calculated over every single transition individually, the resulting set of parameters can be very large. (Since each of these parameters are independent measurements, we denote them as separate scalar values.) Our next step is to constrain the model by requiring that different types of units should share the same behaviour for at least their stable points.

C. Reference values

In the unconstrained model, every single transition can be modelled separately. When we start constraining the model, we require that related transitions share the same parameters at their stable points. Specifically, we require that the ref_{start} and ref_{end} values be exactly the same for all clustered transitions, even though the timing values (f_{start}, f_{end}) may be quite different per transition. We refer to these constrained values as *reference values*, and calculate them at different contextual levels.

In our approach to reference-value estimation, we distinguish between two main types of reference values: (1) static and (2) dynamic. For static reference values we calculate the mean of the normalised feature vectors over all of the specific phone units in the training corpus. We utilise these values to test our initial models.

Dynamic reference values are estimated after a first iteration of trajectory modelling. Once trajectories have been fitted to all transitions, we calculate the reference value means over only those frames associated with trajectory stable points. Since the essence of the initial trajectory model is to identify where the acoustic change takes place, the associated trajectory stable point values serve as a more accurate approximation of the reference values. This is also closer to the behaviour of a traditional HMM, which estimates parameters based on separate states (with a 3-state HMM modelling the approximate left, middle and right of a unit).

Reference values are not only calculated at different contextual levels, but can also be calculated for different groupings of units. For example, all nasal-to-vowel transitions can be grouped together and a single reference value calculated. In the same way, the clusters obtained during triphone tying can also be analysed in a grouped structure.

D. Evaluation and analysis

From the main parameters listed in section III-B, various other values can be calculated. Examples include the slope of transition (the gradient of the first order line), the duration of the slope ($f_{end} - f_{start}$), and the size of the transition ($ref_{end} - ref_{start}$).

In this work, we are interested in determining how well different approaches to trajectory estimation compare with respect to the actual seen MFCC feature vectors of specific classes. In order to do so, the MSE measurement (MSE_{trans}) of the trajectories is particularly useful. This value represents a direct comparison of the model and the training data. The MSE_{trans} measurement can be calculated as follows:

$$MSE_{trans} = \frac{1}{\sum_{s=1}^{S} CF_s} \sum_{s=1}^{S} \sum_{c=1}^{C} \sum_{f=1}^{F_s} SE_{fcs} \quad (4)$$

where SE_{fcs} is the squared error for a specific frame f, a specific coefficient c and a specific sample s.

Every transition generates F squared errors (one for every frame) and there are $C = 13$ of these SE parameter streams (one for every MFCC coefficient stream). To analyse the parameters for all of the examples (S) of a given class, the mean and standard deviation is calculated for the binned trajectories of the same MFCC coefficients.

Lastly, to represent the entire set of transitions with a single error value, we simply sum the contributions from each class:

$$MSE_{global} = \frac{1}{T} \sum_{t=1}^{T} MSE_{trans}, \quad (5)$$

where MSE_{trans} are the mean trajectory MSE estimated for S examples of a contextual class and a total of T classes.

IV. Experimental set-up

A. Overview

For the experiments reported on in this paper, we analyse all of the phone transitions for a set of speech data from a single speaker. As a first step, we track the trajectories for all the training data and estimate the initial approximations. Importantly, these initial trajectories provide us with a timing value (at the start and end point of every transition) as described in section III-A. We use these trajectory alignments for all subsequent trajectory estimations with reference values of different context sizes (only dynamic reference values are discussed here).

B. Speech data

The speech data we use was collected specifically for this analysis in order to provide a large corpus of high quality speech of a single speaker. Only considering a single speaker allows us to focus on contextual effects first, without other speaker specific differences contaminating the results. Based on existing balanced prompt lists [12], we recorded about 2000 short Afrikaans prompts (between 1 to 5 words in length) of a male speaker. (The use of balanced prompt lists ensures sufficient contextual variation.)

A combination of automated and manual review resulted in low quality audio being discarded, and the selection of 1758 of these utterances for analysis. At a total duration of approximately 1.5 hours this corpus produces a higher triphone coverage for a single speaker than is typically available from ASR corpora.

C. Speech segmentation

Our trajectory analysis relies on the identification of accurate phone transition boundaries. We obtain automatic alignments using a standard HMM-based ASR system trained using all 1758 utterances of training data. For this purpose we build a context-dependent cross-word phone recogniser using tied triphone models. 39 MFCC features are used, which include 13 MFCCs and their first and second derivatives. MFCC parameters are computed across a window size of 25ms and a frame rate of 10ms is employed. Each triphone model has 3 emitting states with 7 Gaussian mixtures per state and a diagonal covariance matrix. Cepstral Mean Normalisation (CMN) and semitied transforms are applied. Using a flat-phone grammar, 10-fold cross-validation yields a mean phone accuracy of 90.8%. A forced alignment is performed to output triphone model alignments. The model alignment labels are then converted to the base phone label sequence (the actual phonemes observed in the training data) and used together with the timing information obtained from the alignment to provide the HMM-based phone transition boundaries for speech segmentation.

D. Features for trajectory tracking

Once the transition boundaries have been obtained, MFCCs are extracted for trajectory tracking. These are similar to the ones used during segmentation, except that (1) a 5ms frame rate is used to provide better time resolution, (2) only the raw 13 MFCC coefficients are used and not any derivatives, and (3) the MFCCs are normalised to have zero mean and unit variance. (For each feature vector, normalisation is performed by subtracting the mean and dividing by the standard deviation of the unprocessed feature values.)

Incorporating the phone boundary alignments from above, we associate each of the generated feature vectors with corresponding contextual labelling at the triphone level.

E. Identity-based clustering

Once the transitions have been obtained and labelled, reference values can be calculated by grouping units in different ways, and the corresponding MSEs calculated. We experiment with the grouping of phones based solely on their identities: combining all monophones, combining all biphones and combining all triphones in three different experiments.

To place these experiments in perspective, we also analyse the triphone clusters obtained through acoustic clustering, as described below.

F. Tree-based clustering

During ASR system training, tree-based clustering is performed on a state level using phonetic trees. The system described in section IV-C performs state clustering for each of the emitting states of a specific triphone.

At every node a binary decision is taken based on a context-specific question. For our purposes we include all left and right phone contexts as possible questions. The specific question that is then chosen locally during tree building, maximises the likelihood of the training data given the final set of state tyings [13]. Depending on the answers to these questions a pool of states is successively split. Obtaining meaningful clusters is accomplished using two standard thresholds: (1) Minimum log likelihood TB and (2) the occupation count RO.

The clusters we analyse have been created using optimized thresholds. We perform flat-phone recognition for the single mixture models after tree-based clustering and select values $RO = 24$ and $TB = 80$.

(A good balance for the influence of both parameters is obtained at this point.)

After the tree-based clustering step a set of tied models is generated. The HMM-model structure allows each 3-state triphone model to be identified with unique context labels. We analyse the model structure and perform a look-up on all of the seen triphone labels. It is then possible to track the state-specific model assignments trough the state-specific group labels, previously assigned during tree-based clustering. We use these same labels as the cluster names. Lastly it is also necessary to consider the cluster state (since state clustering is performed for each of the emitting states of a base phone separately). We select only the cluster labels corresponding to the first and last emitting states and determine the pool of triphone labels associated with each of these clusters.

V. Results

In order to determine the overall effectiveness of the trajectory tracking technique, we calculate the MSE_{global} values for different options of reference values. Specifically we analyse (1) the overall accuracy of approximation, (2) the accuracy of approximation for specific transitions, and (3) trends observed for broad transitional classes.

A. Overall accuracy

Reference value	μ	σ
No ref	0.110	0.036
Triphone	0.339	0.073
Triphone Tree-based	0.392	0.097
Biphone	0.433	0.089
Monophone	0.503	0.117

TABLE I

Overall MSE_{global} measurements for different trajectory estimation options calculated over the total number of 38001 *phone transitions*

The values of Table I show the mean μ and standard deviation σ, when the 38001 MSE_{trans} measurement values (one for every phone transition) is reduced to a single value by taking the mean (MSE_{global}) as described in section III-D.

The first value presented ('No ref') lists the error observed if stable points are allowed to be estimated in an unconstrained fashion. This produces a very low overall MSE_{global}, which indicates that the piecewise linear approximation used for transition modelling, is on average effective in capturing trajectory behaviour.

The replacement of the stable points with dynamic reference values at the triphone level yields the second most accurate representation of the training data. However, generalising over trajectories does come at a price. Here the additional error increase is roughly as much when moving from trajectory-specific to triphone-reference values as stable points, compared to only moving from triphone to monophone reference value levels.

As expected, trajectories estimated using the tree-based reference values outperform the biphone reference values. Interestingly, though, the tree-based values seem to be much worse than triphone error values (they are closer to biphones), when we have sufficient triphone examples. We now investigate this effect further.

B. Transition accuracy

The same general measurement can be performed for specific transitional classes. We obtain the MSE_{trans} measurement of section III-D for the examples of a specific phone transition. For simplicity, we also consider only monophone transitional classes (diphones).

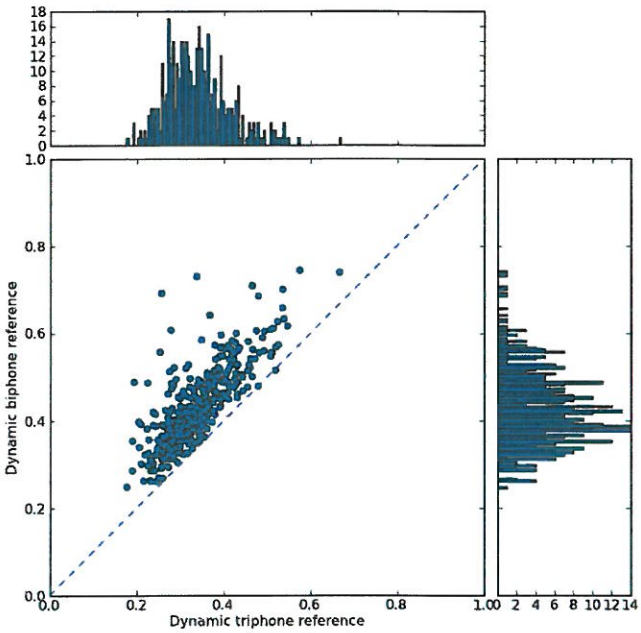

Fig. 1. Comparing trajectory tracking using MSE_{trans} value measurements for the same phone transitional classes of *triphone* and *biphone* reference values

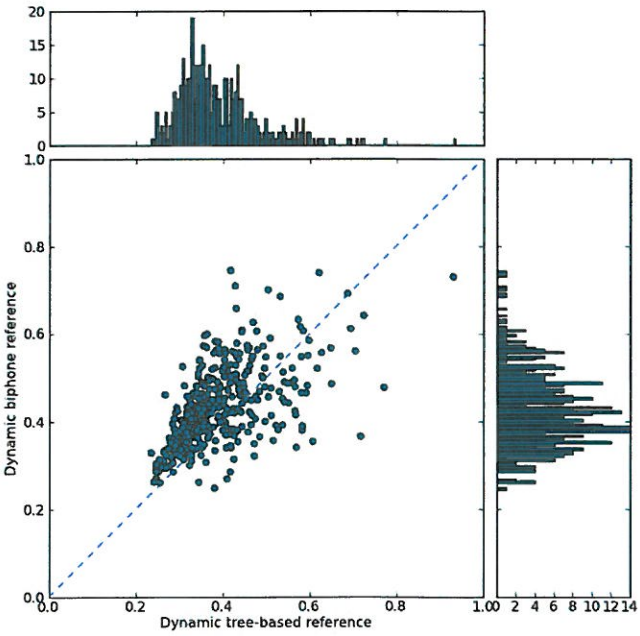

Fig. 2. Comparing trajectory tracking using MSE_{trans} value measurements for the same phone transitional classes of *tree-based* and *biphone* reference values

(Note that our selection of reference values for biphone and triphone trajectory estimation does take into account the phone context of a particular phone transition to make adequate selections.)

Performing these measurements, we generate a single MSE_{trans}

value per transitional phone class. Figure 1 shows a scatter plot of these values when comparing transitional phone classes for triphone and biphone reference values. (Only transitional classes with at least 10 phone examples are selected, to ensure adequate estimation.) The scatter plots help us to (1) compare the specific phone transitional classes on a one-to-one basis and (2) observe the amount of correlation (shape) of these comparisons. It is apparent that all transitions for triphone reference values fit the speech data better than the same transition using biphone reference values for trajectory estimation. We also observe a correlation ($\rho = 0.783$) between the biphone and triphone data points. Importantly, this indicates that the influence of broader phonetic effects is preserved when moving from the triphone to the biphone level.

In contrast to Figure 1, Figure 2 shows the scatter plot when comparing the same transitional classes, but for biphone and tree-based reference values. The correlation drops significantly ($\rho = 0.571$), indicating that less of the broader phonetic effects seen between the triphone and biphone cases are preserved. Furthermore, while the overall measurements for the tree-based reference values outperform the biphone level (see also Table I), this is clearly not the case for all phone transitional classes.

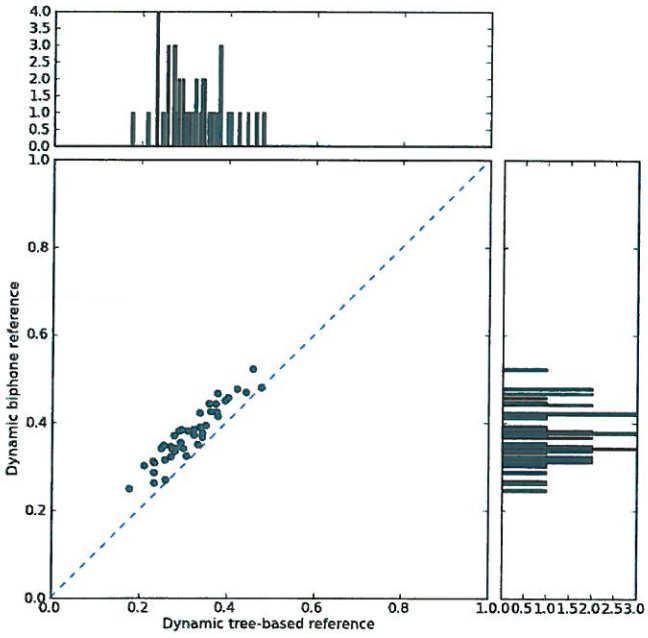

Fig. 3. Good correlation between *biphone* and *triphone* reference values for transitions where *biphone* reference values outperform *tree-based* reference values when tracking trajectories using MSE_{trans} value measurements

In Figures 3 and 4 we report specifically on the biphone values that outperform the tree-based reference values for trajectory estimation. Only transitional classes are selected, where the difference between triphone and biphone reference values are small (For the purposes of this paper we calculate the difference between the global measurements for triphones and biphones and use this value as a threshold). As a last constraint the biphone error value also has to be less than that of the tree-based reference trajectories (Table II).

We observe that the biphone values which are closest to the triphone values and improve on the tree-based errors are highly correlated, showing strong relationships for the within-class error for

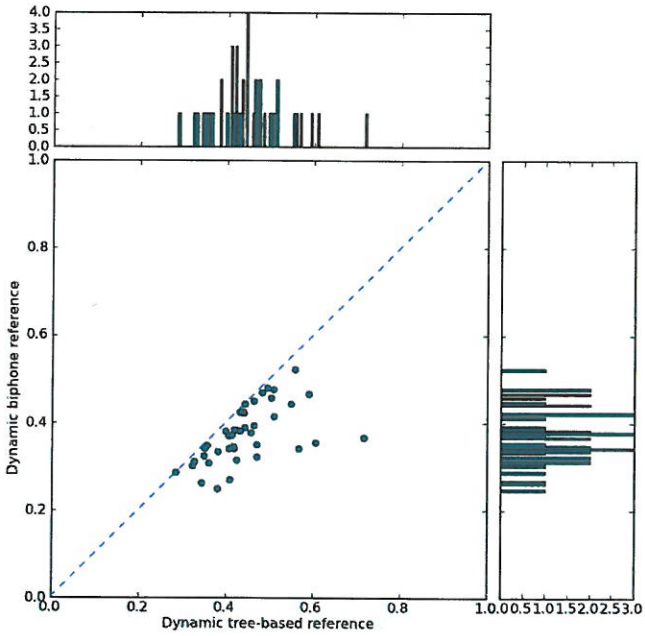

Fig. 4. Additional mismatch when comparing *biphone* and *tree-based* reference values for transitions where *biphone* reference values outperform *tree-based* reference values for trajectory tracking using MSE_{trans} value measurements

Broad class	Trans	Tri	Bi	Tree	Num Frames
vowels-diphthongs	y_i@	0.178	0.248	0.381	2275
nasals-vowels	m_u:	0.211	0.301	0.324	2171
vowels-approximants	@_j	0.231	0.311	0.328	4901
vowels-fricatives	i_v	0.233	0.285	0.286	5122
vowels-diphthongs	i_i@	0.234	0.262	0.345	9334
vowels-fricatives	y_s	0.234	0.307	0.361	3952
...	...	...	...	...	...

TABLE II
Transitions where biphone *reference values perform better than* tree-based *reference values (Transition phone labels in SAMPA)*

these cases. When these same transitional classes are shown for the biphone comparison with the tree-based errors (Figure 4), it is evident that the tree-based reference values used for trajectory estimation, introduce additional mismatch for these transitional classes in our data set.

C. Broad class comparison

Broad class	μ_{Tri}	μ_{Bi}	μ_{Tree}	μ_{Mono}	Num Frames
fricatives-*	0.325	0.390	0.347	0.426	1,813,487
nasals-*	0.327	0.396	0.370	0.429	1,011,660
vowels-*	0.328	0.394	0.354	0.452	3,239,301
approximants-*	0.340	0.398	0.368	0.485	498,576
diphthongs-*	0.371	0.470	0.371	0.506	667,082
trill-*	0.371	0.457	0.384	0.529	715,169
stops-*	0.392	0.459	0.405	0.491	1,895,881

TABLE III
Overall MSE_{trans} measurements for different trajectory estimation options calculated for broad phone clases

Broad class	μ_{Tri}	μ_{Bi}	μ_{Tree}	μ_{Mono}	Num Frames
$*$-nasals	0.327	0.399	0.344	0.447	870,389
$*$-fricatives	0.334	0.405	0.349	0.444	1,700,712
$*$-stops	0.340	0.428	0.361	0.472	1,826,435
$*$-vowels	0.353	0.430	0.387	0.466	2,851,563
$*$-trill	0.357	0.433	0.363	0.497	608,036
$*$-approximants	0.357	0.448	0.409	0.566	425,802
$*$-diphthongs	0.391	0.463	0.388	0.536	627,523

TABLE IV
Overall MSE_{trans} measurements for different trajectory estimation options calculated for broad phone clases

To further understand the strengths and weaknesses of the various context-modelling approaches we compute the MSE_{trans} values when grouping phone classes together. All transitions *from* a specific broad class and all transitions leading *to* a specific broad class are considered. These results are shown in Tables III and IV respectively. The columns *Num Frames* denote the total number of distinct frames summed over when all broad class transitions matching the given specification is accumulated.

In both cases, the strong correlations for the triphone and biphone reference values are immediately apparent (ordering with regard to μ_{Tri} leads to good orderings of μ_{Bi}). We also see some ordering for the tree-based values, but these correlations are not as strong as for the other groups.

VI. Conclusion

In this paper, a piecewise linear trajectory model was presented that is able to model phone transition behaviour at the ceptral level. By comparing variations of the model that group certain units together, the implications of modelling at different contextual sizes can be better understood.

Of specific interest is the extent to which a simple linear model can model phone transitions, as well as the large relative discrepancy between unconstrained and triphone models (given that the analysis is still performed for a single speaker only). Comparatively, the discrepancy between biphone and triphone models is significantly less. We found the triphone models always outperform the biphone models for this analysis. However, different co-articulation effects as shown in [11] can be expected to have varying degrees of influence for the application of the technique on the context level. It would be interesting to consider the very definite phone transitional cases and see whether these clases are indeed closer for biphone and triphone models. Also of interest is that systematic differences in error at various contextual levels can partially be traced back to broad phone categories.

Finally, we found that traditional tree clustering as used for the typical HMM-systems is not always a good context model: a simple biphone model is sometimes a more accurate representation of the acoustics of the unit, despite having fewer free parameters. Future work will build on this finding to determine whether trajectory information can be used to better predict the behaviour of rare or unseen triphones for ASR modelling purposes.

References

[1] K.-F. Lee, "Large-vocabulary speaker-independent continuous speech recognition: The sphinx system," Ph.D. dissertation, Carnegie Mellon University, 1988.

[2] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *In Proceedings of ARPA Workshop on Human Language Technology*, Plainsboro, March 1994, pp. 307–312.

[3] M. Gales and S. Young, "The application of hidden markov models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2007.

[4] H.-A. Chang and J. R. Glass, "A back-off discriminative acoustic model for automatic speech recognition," in *Proc. Interspeech*, September 2009, pp. 232–235.

[5] T. Ko and B. Mak, "A fully automated derivation of state-based eigentriphones for triphone modeling with no tied states using regularization," in *Proc. Interspeech*, August 2011, pp. 781–784.

[6] L. Deng, G. Ramsay, and D. Sun, "Production models as a structural basis for automatic speech recognition," *Speech Communication*, vol. 33, no. 2-3, pp. 93–111, 1997.

[7] D. Yu, L. Deng, and A. Acero, "A lattice search technique for a long-contextual-span hidden trajectory model of speech," *Speech Communication*, vol. 48, no. 9, pp. 1214–1226, 2006.

[8] K. C. Sim and M. J. F. Gales, "Discriminative semi-parametric trajectory model for speech recognition," *Computer Speech and Language*, vol. 21, no. 4, pp. 669–687, October 2007.

[9] K. Tokuda, H. Zen, and T. Kitamura, "Trajectory modeling based on HMMs with the explicit relationship between stochastic and dynamic features," in *Proc. Eurospeech*, September 2003, pp. 865–868.

[10] M. Ostendorf, V. Digalakis, and O. Kimball, "From HMMs to segment models: A unified view of stochastic speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 4, no. 5, pp. 360–378, May 1996.

[11] J. A. C. Badenhorst, M. H. Davel, and E. Barnard, "Analysing co-articulation using frame-based feature trajectories," in *Proc. PRASA*, November 2010, pp. 13–18.

[12] N. J. de Vries, J. Badenhorst, M. Davel, E. Barnard, and A. de Waal, "Woefzela - an open-source platform for ASR data collection in the developing world," in *Proc. Interspeech*, August 2011, pp. 3177–3180.

[13] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Veltchev, and P. Woodland, *The HTK Book*. http://htk.eng.cam.ac.uk/: Cambridge University Engineering Department, 2005.