Analysing co-articulation using frame-based feature trajectories

Jaco Badenhorst
[1]School of Electrical, Electronic and
Computer Engineering, North-West University,
Potchefstroom, South Africa
[2]Human Language Technology
Competency Area,
CSIR Meraka Institute
Email: jbadenhorst@csir.co.za

Marelie H. Davel
Human Language Technology
Competency Area,
CSIR Meraka Institute
Email: mdavel@csir.co.za

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

***Abstract*—We investigate several approaches aimed at a more detailed understanding of co-articulation in spoken utterances. We find that the Euclidean difference between instantaneous frame-based feature values and the mean values of these features are most useful for these purposes, and that low-order polynomials are able to model the between-phone transitions accurately. Examples of typical transitions are presented, and shown to give useful insights on the measurable effects of co-articulation.**

I. Introduction

With current technology, it is generally agreed that large amounts of training data are required to achieve high accuracies in speech-recognition systems: state-of-the-art large-vocabulary systems are trained with hundreds to thousands of hours of data. However, it is not clear just why so much data is required: is it because of inherent variability in speakers, channel conditions, speaking styles, etc., or because of the complexity in representing cross-phone co-articulation accurately, or for some other reason? This issue is theoretically important, and also crucial for the development of systems in resource-constrained environments.

An interesting hint on this matter is provided by the performance of typical Hidden Markov Model (HMM) systems on different sub-corpora of the TIMIT corpus [1]. In particular, we have repeatedly found that performance is substantially better on the so-called speaker-independent sentences (the sa subset, where the same prompts are recorded by all training and testing speakers) compared with the speaker-dependent sentences (the si subset, where different prompts are recorded for different speakers, and each of the sentences is thus only recorded once). In Table I we list phone recognition accuracies obtained for subsections of the testing data containing the indicated sentences. All accuracies are obtained using the same HMMs, constructed from the training set. (Table I also contains the results for the sx sentences, which were read by small subsets of the speakers – these clearly behave similarly to the speaker-dependent sentences.)

Since these sub-corpora are subjected to the same intra- and inter-speaker sources of variability, the large accuracy difference between the sa sentences and the other two sentence types suggests that context modelling (and thus co-articulation) plays a significant role in the accuracy of speech-recognition systems – thus, also in their need for large training corpora. It is clearly not enough to see a sufficient number of phone samples: it is necessary to see enough samples in contexts sufficiently similar to what is observed in the testing data.

Subset	Gender	% Accuracy
sa	male	88.78
sa	female	87.47
si	male	61.24
sx	male	61.13
sx	female	57.46
si	female	56.20
Total	-	65.28

TABLE I
Typical accuracies of different sentences in TIMIT test data set

We would like to gain a more detailed understanding of these contextual effects. Towards this goal, we have developed a number of tools that allow us to assess how phonemic context influences the production of speech sounds, when expressed in terms of the standard features used for speech recognition. In this paper we introduce these techniques, and demonstrate their usefulness in analysing co-articulation effects.

The paper is structured as follows: We first discuss some related research in section II. In section III we describe the specific techniques we use to analyse contextual effects. We then describe the experimental set-up that we use to test the validity of these techniques and to perform initial experiments in section IV. Our results are presented in section V, followed by a summary of our main observations and a preview of future work, in section VI.

II. Background

While the importance of modelling contextual effects for large vocabulary speech recognition has long been understood [2], these effects are typically modelled implicitly within a more general statistical framework. Attempts to model contextual effects explicitly as phone transition trajectories have been met with mixed success [3]. Most of these approaches attempt to overcome the limitations of standard HMM approaches (especially the state-based independence assumption) either by incorporating explicit trajectories within an HMM framework [4] or by explicitly defining longer term variable length segmental models [5]. Related research tries to uncover the underlying articulatory trajectories producing speech, in an attempt to better model acoustic change with fewer parameters [6].

All the above approaches aim to develop better acoustic models of speech. Much less work is available related to an analysis of co-articulation effects as a tool towards a better understanding of speech resource requirements, the focus of the current paper.

III. Technique

In this section we describe our analysis technique in general, list some of the parameters that can be varied, and discuss the design choices made.

The essence of the analysis technique is to identify reference values per phone, and then track the trajectory with which the audio signal diverges from these reference values over time. We expect these reference values to act as if they are 'targets', with some form of transition occurring from one target to the other over time. We are interested in determining whether different types of transitions occur, and whether similar transitions are observed over similar phone classes across multiple speakers.

A. Reference values

Typical ASR systems utilize frame-based feature vectors such as Mel-Frequency Cepstral Coefficients (MFCCs) or Linear Predictive Coding (LPC) coefficients to represent the speech signal effectively. In this work we utilise MFCCs normalised to have zero mean and unit variance as our input features. (For each feature vector, normalisation is performed by subtracting the mean and dividing by the standard deviation of the unprocessed feature values.) All MFCC vectors are generated using the same parameters as the system described in section IV-B.

As reference values, we calculate the mean of the normalised feature vectors over all monophones in the training corpus. ASR alignments are always used to associate the feature vectors with corresponding phone labels, leading to a selection of feature sections that would normally be selected during the ASR training process. Different means can be calculated by either summing over all speakers or only over monophones of the specific speaker. In addition, all frames in a monophone can be used, or only the central frames (associated with the centre states of the HMM alignments, assumed to be more stable as target values, and less subject to co-articulatory effects).

B. Difference measures

Various analytical functions may be used to calculate the extent in which each frame diverges from the respective reference values. We experiment with the Pearson correlation coefficient, the Euclidean distance, and the dot product between two vectors.

These measures are defined as follows. For any two random variables X and Y the Pearson correlation coefficient is given by:

$$\begin{aligned} \rho_{XY} &= \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \qquad (1) \\ \text{where } Cov(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \end{aligned}$$

and $\mu_.$ and $\sigma_.$ indicate the mean and standard deviation of each of the variables. The Euclidean distance is given by:

$$d_{XY} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (2)$$

where x_i and y_i are the separate dimensions of the n-dimensional random variables X and Y, and the dot product by:

$$X \cdot Y = \sum_{i=1}^{n} x_i y_i \qquad (3)$$

C. Tracking trajectories

Each of the above measurements are used to obtain two discrete values per frame (measuring the difference from the two reference values on either side of the transition boundary). In order to create a trajectory from the frame-based values, we fit a polynomial function using least-squares estimation. This approach effectively minimizes the squared error E, given by:

$$E = \sum_{i=1}^{n} |p(x_i) - y_i|^2 \qquad (4)$$

where $|p(x_i) - y_i|^2$ are the squared residuals.

The order of the polynomial is an important factor to consider, with higher order polynomials quickly leading to overfitting. We describe the trajectories formed near transition boundaries in terms of a 3rd order polynomial function and only fit the frame sequence closest to the phone boundary. This is done in order to prevent interference from additional co-articulation to the left and right of the phone transition being analysed. (Only the closest 50% of monophone frames are used in our experiments, effectively describing a diphone, a heuristic measure meant to obtain a balance between including only the relevant part of the trajectory and still retaining sufficient frames for analysis.)

In order to model a phone transition, two trajectories – one using each reference variable as target – are constructed.

D. Measuring co-articulation effects

In order to analyse co-articulation effects, we measure:

- the goodness of fit per trajectory,
- the difference between monophone reference values, and
- the trajectory slope at the transition boundary.

We analyse these measurements over all test data, and for specific phone classes. We also report on the standard deviation of these measures as an indication of intra-class variability.

IV. Experimental set-up

A. Overview

Frame-based values can be calculated using any of the difference measures and reference mean variables described above. In order to ensure that we are constructing meaningful representations of the modelled acoustics, assessment of the difference measures are required. In essence, given specific reference mean variables (corresponding to the phone labels of a transition) tracked trajectories must yield the best possible separability of the frame-based features to the left and right of a transition boundary. (Some transition classes have such strong co-articulation effects that separation is acoustically constrained. This will typically be the case for very similar sounding phones.) In our first set of experiments, reported on in section V-A, we use class separability and boundary tracking to evaluate the overall accuracy of our technique.

Extraction of meaningful trajectories from the frame-based values is achieved using polynomial functions. The different ways in which these trajectories categorise different types of acoustic change is investigated in our next set of experiments, reported on in section V-B.

Co-articulation effects manifest differently for different phone contexts. To understand how co-articulation phenomena can be analysed based on the constructed trajectories, we conduct experiments considering broad phone transition classes, as reported on in the final part of the results section (section V-C).

B. Speech data and alignments

We use the TIMIT speech corpus [1] for all of the experiments discussed below. The corpus consists of 630 speakers from eight major dialect regions in the United States. For every speaker there are 10 utterances resulting in a total number of 6300 utterances. The corpus is divided into a standard training and testing set. For the training data there are 326 male and 136 female speakers giving a total of 462 training speakers. The types of sentences that were read is divided into three parts: sx, si and sa. MIT designed the 450 phonetically balanced sx sentences, while the si sentences form 1890 phonetically diverse sentences designed at SRI. Finally the test set consist of 168 speakers, selected so that no speaker appears in both the training and test set.

In order to generate accurate phone transition boundaries, we obtain automatic alignments using a standard HMM-based ASR system trained using the training set of the TIMIT corpus. 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 order derivatives. MFCC parameters include a window size of 25ms and a frame rate of 10ms respectively. Cepstral Mean Normalisation (CMN) is applied. With regard to the modelling structure, each triphone model has 3 emitting states with 7 Gaussian mixtures per state and a diagonal covariance matrix. The system is used in forced alignment mode to output state-level phone alignments. These alignments provide the HMM-based phone transition boundaries used in the next section.

V. Results

A. Overall accuracy of measures

In order to evaluate the effectiveness of the trajectory tracking technique, the frame-based values are analysed with regard to: (1) their ability to separate classes to the left and right of the known phone transition boundary, and (2) the proximity of the trajectory-based transition boundary to the HMM-based transition boundary.

1) Class separability: It is possible to measure the average difference from each reference value (the average of the frame-based values) to the left and right of the (known) transition boundary, and perform phone classification based on the difference between these two values. Table II indicates the number of phone transitions for which both of the phones are correctly classified using the various difference measures described in section III-B. It is found that, while all three difference measures provide fair class separability, the Euclidean distance outperforms both correlation and the dot product.

Switching to the state level boundaries (indicated as *ASR centre* in the table) results in an even further improvement for the Euclidean distance, but not for the other measures. This shows the presence of two opposing effects: (1) stationary components at phone centres and (2) encoding of co-articulation in the reference variables. At the phone centres less co-articulation yields more separable trajectories, while longer trajectories are likely to reveal more information with regard to the particular phone.

The observed classification accuracy (averaged over all phone classes) of 81.3% when using the Euclidean distance as difference measure is surprisingly high, given the simplicity of the classification technique. For the remainder of our analysis, we mainly report on results obtained using the Euclidean distance. Similarly, we focus on the use of speaker-independent monophone means as reference values. Calculating a complete set of classification results, given the different options of reference values, yield only slightly better classification for speaker-specific means or means based only on central frames. We find the speaker-independent monophone means more robust because of the large amount of data available in the training corpus.

Difference measure	# Correct classifications	% Accuracy
Euclidean	40 558	77.1
Correlation	39 190	74.5
Dot product	36 644	69.7
Euclidean (ASR centre)	42 747	81.3
Correlation (ASR centre)	37 585	71.5
Dot product (ASR centre)	31 074	59.1

TABLE II

Number of correct classifications using mean frame-based values and known ASR boundaries

2) Boundary tracking: The evaluation technique described above relies on a known transition boundary. How close is the transition boundary identified by the tracked trajectories from the version obtained from the HMM-based ASR system? We evaluate this for different orders of polynomial functions, using the crossing points of the two trajectories to identify transition boundaries.

Not all phone transitions produce pairs of trajectories that cross each other: Table III lists the number of phone transitions that can be identified using polynomial function crossing points. For the usable boundaries, the distance (in frames) between the identified and known phone transition boundaries is calculated. This provides a clear indication of the boundary tracking capability of these functions. (Note that the ASR-based boundaries are also estimates rather than an indication of a ground truth.)

In Table III we also report on the goodness-of-fit (E) for the different polynomial functions, calculated by taking the average of the mean square error values that describe the fit of the two individual polynomial functions. As higher order functions are used to estimate trajectories, a closer fit is obtained and the mean square error decreases. As this may lead to overfitting, we select a 3rd order polynomial for the remainder of our analysis: the shape of a 3rd order polynomial lends itself well to describe the behaviour of a trajectory near and crossing a phone boundary, and allows us to focus on the co-articulation due to a single phone transition.

Measure (order)	# Usable boundaries	% Usable boundaries	E	Diff (# frames)
Euclidean (1)	42 552	80.9	6.345	1.828
Euclidean (2)	47 909	91.1	4.318	2.100
Euclidean (3)	48 737	92.7	2.990	1.897
Euclidean (4)	49 312	93.8	2.311	1.846
Correlation (1)	41 502	78.9	0.964	1.839
Correlation (2)	46 430	88.3	0.569	2.103
Correlation (3)	47 534	90.4	0.339	1.933
Correlation (4)	48 101	91.5	0.240	1.871
Dot product (1)	39 526	75.2	44.471	1.959
Dot product (2)	46 079	87.6	25.051	2.314
Dot product (3)	46 688	88.8	13.976	2.039
Dot product (4)	47 272	89.9	9.580	1.971

TABLE III
Boundary tracking of phone transitions using different orders of polynomial functions

B. Trajectory models

From the results in Section V-A it can be seen that the underlying speech features (MFCCs) are co-articulated in two main ways: 1) Dynamics of change characteristics 2) acoustic contextual influence. These two effects may also interact with each other.

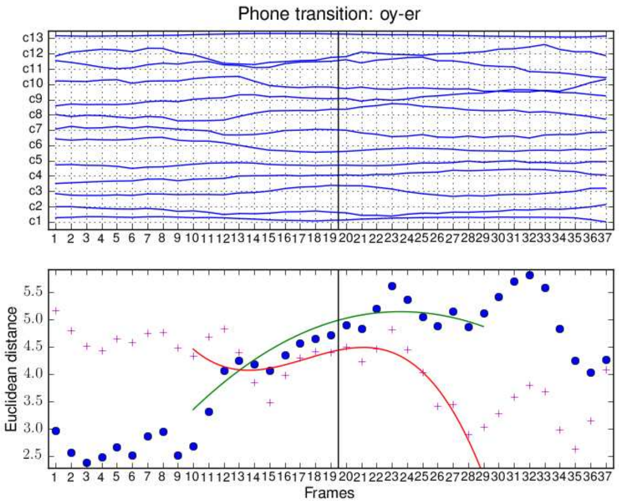

Fig. 1. Gradual trajectories revealing strong co-articulation for the vowel-vowel phone transition.

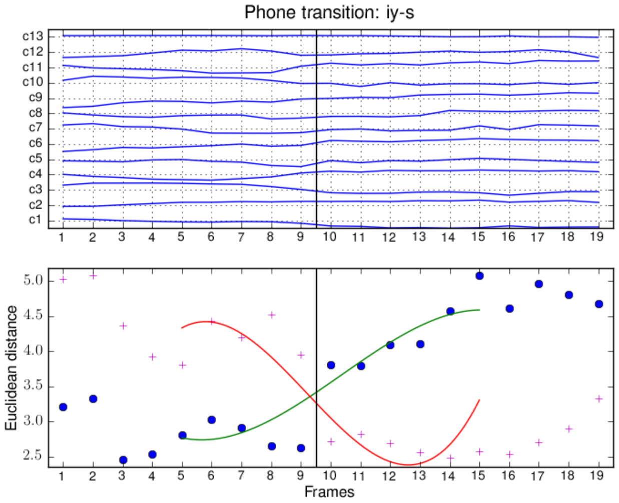

Fig. 2. Steep trajectory slopes revealing the definite transition of the vowel-fricative class

To show the prominent types of co-articulations observed, we present four example figures. The plots show the stacked 13 MFCC coefficients for all frames of the monophone transition, the Euclidean frame-based difference values, as well as the final diphone trajectories consisting of the two polynomial functions. Blue dots indicate frame-based values for the first phone, similarly red crosses correspond to frame-based values for the second phone label and the phone transition boundary as identified by the HMM-based ASR system is indicated as a vertical line.

Figure 1 represents an example of the phone transition /oy/-/er/ within the vowel-vowel class, spoken by a male. Strong acoustic co-articulation over a relatively long period of time is clearly visible for frames 11 - 27. This results in a gradual change and small slope values at the ASR boundary. From the frame-based values, one can see that classification with regard to the mean value is still possible, assisted by the long duration of the speech segment.

An example of a female /iy/-/s/ transition belonging to the vowel-fricative class, is given in Figure 2. The MFCCs and

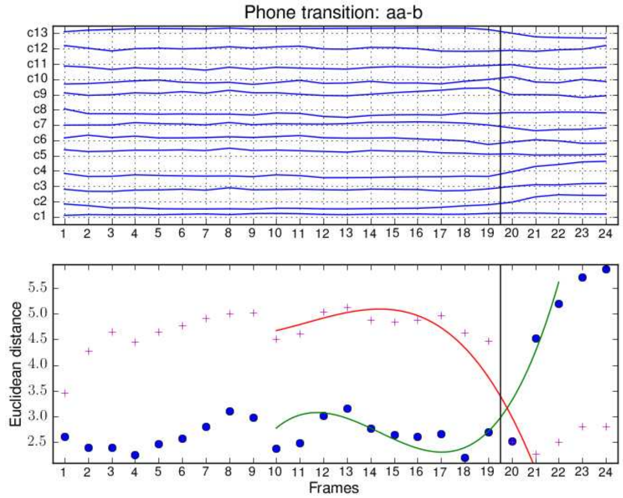

Fig. 3. Abruptly changing trajectories of the vowel-stop phone transition class

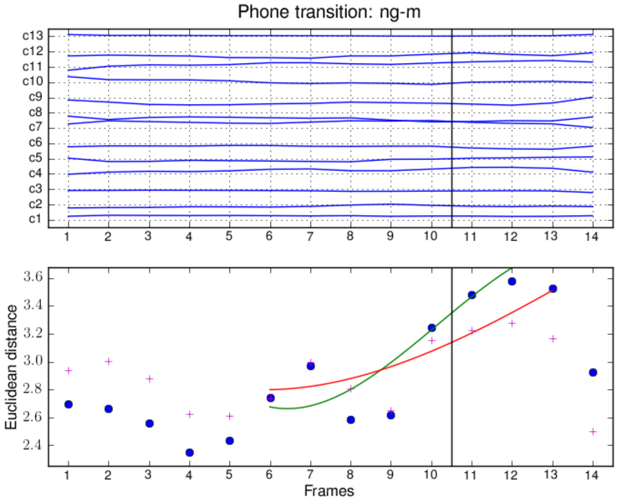

Fig. 4. Low separability and strong co-articulation effects yield similar trajectories for the nasal-nasal class

frame-based difference values clearly show a definite transition around frame numbers 9-10, indicating a large difference in acoustic quality between the two phones. It is interesting to note that even for large acoustic change, co-articulated features flowing well into both phones are present. Diphone polynomial trajectories have steep slopes at the ASR boundary and classification with regard to the average of the frame-based values is accurate.

There are also abrupt transitions, with very little co-articulation visible. A clear example comes from the vowel-stop class (/aa/-/b/). Both MFCCs and frame-based difference values show very fast change within a small time period. Co-articulation effects with regard to this transition is seen to affect only 4 frames 18-22 and the frame-based values have high separability (see Figure 3).

During all of the analyses (also see below) the nasal-nasal transition class tends to be problematic. From the MFCC values shown in Figure 4, the straight lines indicate very similar acoustic quality for most of the frames and only gradual changes. The frame-based difference values support this finding, showing only gradual transition and bad separation. Co-articulation is seen to be present for all of the frames, although this may be influenced by the similarity of the two targets being tracked. The slopes of the polynomials have the same sign and are very similar.

In this section we demonstrated the use of trajectory models to analyse co-articulation by presenting four very specific examples that are prototypical of the types of co-articulation observed in the larger corpus. In the next section we analyse some of these effects by averaging over broad phonetic classes.

C. The effect of broad phonetic classes

Different classes of phone transitions reveal interesting trajectory effects. Specifically, we evaluate 5 parameters to categorise the trajectories formed for different classes:

1) The slope of the polynomial function trajectory of the first phone reference variable,
2) the slope of the polynomial function trajectory for the second phone reference variable,
3) the Euclidean distance of the monophone means (the difference between the two reference values),
4) the standard deviation σ_1 of the first slope, and
5) the standard deviation σ_2 of the second slope.

Table V shows the above values for different phone transition classes constructed according to the CMU dictionary phone groupings [7].

Ordering with regard to the steepness of the slopes, we see that phone transitions with steep slopes also yield good separation for the mean difference between the reference values. Indeed we calculate the average of the frame-based values for the vowel-fricative, vowel-affricate, nasal-fricative, nasal-affricates, to yield correct classification percentages of 91.6, 95.8, 84.4 and 92.1 respectively (Table IV). Similarly, the nasal-nasal class has a low separation of the average frame-based values (0.194 - see Table V) for the few phone transitions (29.1%) that do yield correct classification. There are exceptions to the rule. The nasal-liquid class has good separability and steep slopes but classification of the average frame-based values is at 60.0%. In general, similar classes (such as nasal-nasal or fricative-fricative) have the weakest separability, as could be expected.

We observe that the standard deviations of the slopes, σ_1 and σ_2 to be similar in magnitude for particular phone classes, indicating a similar variability with respect to the intra-class diphone transitions. Interestingly, the magnitude of the two slopes are typically not equal, with the divergence from the first mean occurring more quickly than the approach towards the second mean. This co-articulation effect warrants further investigation.

Transition group	# Correct classifications	% Accuracy
vowel-affricate	640	95.8
vowel-fricative	8 268	91.6
vowel-semivowel	2 096	83.1
vowel-stop	9 142	79.8
vowel-nasal	5 005	76.5
vowel-vowel	1 143	70.2
vowel-aspirate	693	65.4
vowel-liquid	4 790	62.9
nasal-affricate	93	92.1
nasal-fricative	862	84.4
nasal-semivowel	125	79.6
nasal-aspirate	38	64.4
nasal-stop	1 040	63.1
nasal-liquid	183	60.0
nasal-nasal	23	29.1
liquid-affricate	53	100.0
liquid-fricative	709	94.8
liquid-stop	1 969	83.9
liquid-semivowel	230	79.6
liquid-liquid	44	73.3
liquid-aspirate	30	65.2
fricative-semivowel	353	94.9
fricative-aspirate	70	75.3
fricative-stop	1 864	69.5
fricative-fricative	270	59.3
fricative-affricate	34	57.6
stop-semivowel	408	72.6
stop-affricate	77	63.6
stop-aspirate	52	45.6
stop-stop	227	35.0
semivowel-affricate	9	81.8
semivowel-aspirate	15	68.2
affricate-aspirate	3	100.0
total	40 558	77.1

TABLE IV
Number of correct classifications using mean frame-based values and known ASR boundaries for specific transitions.

VI. Conclusion

It is clear that polynomial models of the Euclidean difference between the mean and instantaneous MFCC vectors are highly informative on the nature of the transitions between different phone classes. These transitions, in turn, capture the essence of the co-articulation effects which – according to the argument in Section I – are likely to be an important factor in the substantial data requirements for high-accuracy speech recognition systems.

In light of the variability seen in the different types of phone transitions, it is not surprising that current context models do not generalize well to unseen (or rarely seen) context-dependent phones. This suggests that models tailored to the different types of transitions seen here may lead to systems that are more parsimonious in their data needs; the development of such models is therefore the major focus of our ongoing research.

Transition group	Slope 1	Slope 2	Diff reference values	σ_1	σ_2
vowel-fricative	0.426	−0.236	3.064	0.510	0.516
vowel-stop	0.427	−0.203	2.925	0.603	0.592
vowel-affricate	0.353	−0.205	2.964	0.432	0.398
vowel-nasal	0.376	−0.164	2.493	0.639	0.635
vowel-semivowel	0.230	−0.180	2.413	0.490	0.515
vowel-vowel	0.168	−0.182	2.561	0.292	0.289
vowel-liquid	0.164	−0.175	2.551	0.349	0.388
vowel-aspirate	0.123	−0.063	2.122	0.440	0.466
nasal-liquid	0.347	−0.290	2.782	0.495	0.441
nasal-fricative	0.295	−0.317	2.819	0.531	0.485
nasal-affricate	0.311	−0.289	2.486	0.303	0.313
nasal-semivowel	0.566	0.001	2.108	1.311	0.924
nasal-stop	0.254	−0.266	1.958	1.030	0.943
nasal-aspirate	0.230	0.008	1.792	0.690	0.896
nasal-nasal	−0.226	−0.394	0.983	1.553	1.746
liquid-fricative	0.499	−0.315	3.084	0.528	0.503
liquid-affricate	0.358	−0.341	3.308	0.642	0.482
liquid-stop	0.371	−0.240	2.898	0.715	0.665
liquid-liquid	0.231	−0.207	2.944	0.269	0.298
liquid-aspirate	0.102	−0.126	2.440	0.436	0.415
liquid-semivowel	0.074	−0.084	3.030	0.241	0.262
fricative-semivowel	0.259	−0.323	3.140	0.429	0.457
fricative-stop	0.142	−0.154	1.760	0.427	0.444
fricative-aspirate	0.253	0.073	2.037	0.424	0.596
fricative-fricative	0.057	−0.068	1.450	0.539	0.490
fricative-affricate	0.088	−0.028	1.510	0.195	0.211
stop-semivowel	0.006	−0.493	2.573	1.403	1.215
stop-affricate	0.185	−0.107	1.630	0.332	0.308
stop-aspirate	0.059	0.260	1.583	1.080	1.300
stop-stop	0.119	−0.051	1.063	0.700	0.573
semivowel-affricate	−0.063	−0.510	2.951	0.556	0.320
semivowel-aspirate	0.321	0.0156	2.119	0.507	0.466
affricate-aspirate	0.503	0.096	1.635	0.301	0.362

TABLE V
Slopes of 3rd order polynomial functions at ASR diphone transition boundary

References

[1] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus, NIST order number PB91-100354," February 1993.

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

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

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

[5] 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, 1996.

[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] "The CMU pronunciation dictionary," 2010, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.