Improved transition models for cepstral trajectories

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 improve on a piece-wise linear model of the trajectories of Mel Frequency Cepstral Coefficients, which are commonly used as features in Automatic Speech Recognition. For this purpose, we have created a very clean single-speaker corpus, which is ideal for the investigation of contextual effects on cepstral trajectories. We show that modelling improvements, such as continuity constraints on parameter values and more flexible transition models, systematically improve the robustness of our trajectory models. However, the parameter estimates remain unexpectedly variable within triphone contexts, suggesting interesting challenges for further exploration.**

I. Introduction

Current approaches to automatic speech recognition (ASR) require large amounts of speech data to achieve high accuracies, since context-dependent modelling of phones is an important feature of these approaches. The requirement for context-dependent modelling results from the physical constraints of the human vocal tract, which results in co-articulation effects during the transition from one phone to the next. Since state-of-the-art ASR systems model speech with piecewise-constant statistical models, observations of the influences of various phonetic contexts on each phone are required to create adequate statistical models of the effects of co-articulation. Hence, sufficient examples are required for each representative context. Unfortunately, this leads to substantial data requirements.

Trajectory modelling approaches [1], [2] have attempted to model temporal information in a more explicit fashion in order to reduce these data requirements. It is clear that the effects of co-articulation are not constrained to the frame level – appropriate models need to operate at the segmental level, and even longer-term effects must be considered. Describing the observed variability on all these levels is a challenging problem. We are specifically interested to know whether systematic phone transition effects may be described more accurately. It is our belief that finding appropriate representations is important to enable more effective parameter sharing, and thus more data-efficient ASR.

With this work we improve on a model that can be used to isolate the key elements that occur in acoustic features during phone-to-phone transitions. We first show that trajectory tracking may be accomplished for a basic model and the ability of the model to predict trajectory behaviour at different context sizes is further evaluated. To better account for additional trajectory behaviour, a more complex model description is developed to characterise the observed variability.

This paper is structured as follows: Related research is discussed in Section II. Specific techniques used to model phone transitions and the measurement strategies thereof are presented in Section III. We then describe our experimental setup in Section IV and details regarding our experiments and results are given in Section V. Our concluding statements are made in Section VI.

II. Background

Accurate modelling of co-articulation effects in speech data has been the main driving force behind the development of large speech recognition corpora [3]. In fact, if unlimited training data were available, it would be more beneficial to model co-articulatory effects using whole word (or even phrasal) units instead of phones as the basic modelling unit, since co-articulation effects are increasingly well modelled by larger contexts. Limited training data, however, forces the use of smaller units. Context-dependent phones are currently widely used to approximate the co-articulation effects for accurate speech recognition [3]. Finding the correct segment size to model the diversity of all co-articulation effects can, however, prove difficult. A key motivating factor for the development of segmental models is the fact that it is possible to exploit acoustic features that are apparent at the segmental and not at the frame level [1], [2].

The hope is that more data-efficient models of co-articulation can be developed in this way, but this is not a straightforward goal to achieve. One problem is that any segmental approach needs to model extra-segmental variability (between different examples of speech segments) as well as intra-segmental variability (within a single example) accurately. The observed variability for segments of variable length may have multiple sources, and their interaction is currently not well understood. Possible origins for these sources include factors such as recording conditions, different speaking styles, phonetic reduction and finally co-articulation.

A wide variety of approaches to the development of segmental models have been proposed, based on several fundamental observations. For example, in [2] the fact that time-normalised phones tend to behave predictably in various phonetic contexts was used to develop a probabilistic trajectory model. Also, the speech production process suggests the influence of underlying articulatory patterns (trajectories) on speech data [4] and more recently, convolutional non-negative matrix factorisation (CNMF) has been used as an approach to discover temporal (sequential) patterns in speech data [5]. CNMF showed a great deal of time warping variation and therefore time-coded NMF (motivated by findings in neuroscience) has been attempted to improve pattern discovery [5].

Attempts to explicitly model temporal effects (trajectories) in speech data have, to date, achieved limited success [6]. Specific limitations of the HMM modelling paradigm, in particular the state-based independence assumption, are addressed in these methods. This is mainly accomplished by either incorporating explicit trajectories within the HMM framework [7] or by defining longer-term variable-length segmental models [8].

In a novel approach to implement a hidden trajectory model, bi-directional filtering of vocal tract resonances (VTR) yields promising results and also enables the implementation of variable-length representation of long-contextual-spanning speech effects) [9]. Conceptually, the opposite approach is to model the trajectories of the features used for speech recognition directly, and in [10], [11] it was found that such models of cepstral features are able to represent co-articulatory phenomena in a way that makes context dependency explicit. The current paper similarly models the cepstral trajectories directly, and demonstrates how more accurate parameter fits can be achieved by using more sophisticated transition models.

III. Approach

The piece-wise linear approximation that we use to track cepstral trajectories effectively captures temporal changes using sub-phone level segments (as opposed to the individual frames) for every phone transition. Applying a search to find variable-length positions for these segments allows us to characterise detailed transitional behaviour and obtain a direct comparison between the modelled trajectories and the actual speech data. By measuring how consistent the tracked changes are, different modelling choices may be compared, leading to new insights regarding cepstral transition behaviour.

A. Cepstral transition models

We model speech data using MFCC features, which are widely used in state-of-the-art speech recognition systems. Near phone transitions, co-articulatory effects on these features have been shown to be highly regular in [10]. In particular, the phones on either side of a transition generally determine a target value (which the trajectory may or may not reach), and the trajectories generally interpolate fairly smoothly between those targets. The authors of [11] utilised this finding, describing individual phone transition behaviour with a simple piece-wise linear approximation model. Their model consisted of three line pieces to fit the cepstral values (frames) of a single MFCC (cepstral transition), using least-squares optimisation. Start and end line segments were constrained to be constant values. We refer to these constant line segments as *stable values* and the remaining central line segments as the *change descriptor*. To find a complete piece-wise linear approximation for any cepstral transition, a search is required to determine the start and ending indices (model alignments) of the change descriptor. Similarly to the method described in [11], the squared error for all line pieces of the cepstral transition model can then be found, yielding a single error value for each approximation. Optimising the squared error enables us to find the best model alignments. In order to compare the different options, the squared errors (SE_f) of each parameter at each instant are estimated, followed by the mean square error (MSE_{model}) across features:

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

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

$$MSE_{model} \quad = \quad \frac{1}{F}\sum_{f=1}^{F} SE_f \tag{2}$$

In [11] an algorithm is described that allows the piece-wise linear model to share contextual information with other (similar) transitions. By constraining the stable values to reference value estimates of different context sizes, the context dependency of these models can be evaluated. Similarly, what constitutes a single “cepstral transition model” can be specified according to context length, phone identities or even broad classes of phones.

B. Model evaluation

Our first priority in modelling phone transitions is to accurately represent speech data. This will then subsequently serve as the enabling factor, so that systematic effects (if they are present) may be identified. In terms of the models described here, we analyse two main criteria to facilitate these goals. The first measurement (model fit) is used to evaluate the ability of a model to track observed trajectories. Secondly we characterise individual cepstral transitions by evaluating:

- The consistency of a measurement across multiple samples of the same transition in a data set.
- The ability of the model to predict parameters of unseen samples (we estimate transition model parameters on a training set and evaluate the error on a separate test set).

1) Model fit: In Figure 1 the linear approximations for the first four cepstra of a single diphone transition example can be seen. The separate model parts of the first cepstral coefficient (MFC 1) can clearly be identified: two stable values (frames $9-15$ and frames $17-21$) and a single change descriptor (frame 16, connecting the stable values). As this is a segmented model, the stable values are anchored to the start and ending frames of the diphone segment (frames 9 and 21 respectively) and do not extend to adjacent transition frames numbers ($1-8$ or $22-26$). For all cepstra a single definite transition is observed near the ASR boundary.

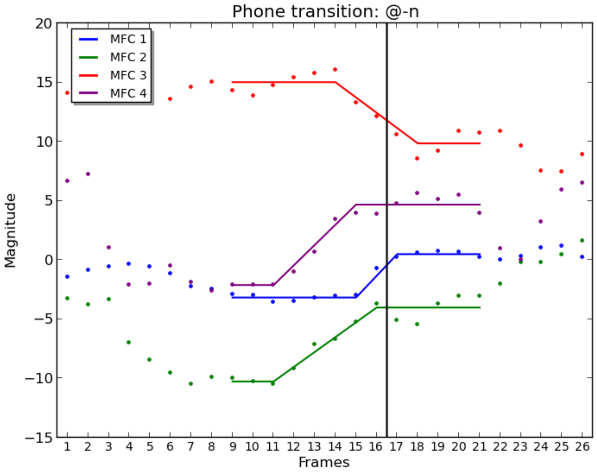

Fig. 1. Piece-wise linear model fit of the first four cepstra of the diphone transition /@-n/ using 3-piece segmented models

Through Equation 2, the MSE measurement can be calculated for the separate model parts, or for the whole piece-wise approximation of the specific coefficient and multiple transition examples, by including the relevant frames. To measure how well different trajectory estimation approaches compare with respect to the actual observed MFCC feature vectors, the MSE measurement (MSE_{trans}) of trajectories is particularly useful. This value allows the direct comparison of phone transitions, with regard to the training data, across all cepstral transition models. The MSE_{trans} measurement can be calculated as

$$MSE_{trans} = \frac{1}{\sum_{s=1}^{S} CF_s} \sum_{s=1}^{S} \sum_{c=1}^{C} \sum_{f=1}^{F_s} SE_{fcs} \qquad (3)$$

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 cepstra (one for every MFCC coefficient). To analyse the parameters for all of the examples (S) of a given class, the mean and standard deviation are calculated for the binned trajectories of the same MFCC coefficients.

Finally, to represent the entire set of transitions with a single error value, the summation of the contributions from each class is evaluated:

$$MSE_{global} = \frac{1}{T} \sum_{t=1}^{T} MSE_{trans}, \qquad (4)$$

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

2) Model consistency: To identify systematic behaviour for cepstral transition trajectories, we present two consistency measurements, in which both stable values and change-descriptors are evaluated. Different modelling options can then be compared directly (for the same transition examples). More consistent model parameters are a more favourable choice for the representation of the transition model.

Reference stable values are estimated using the training data set. These values are obtained in a similar way to that described in [11]. Once an initial set of trajectories have been fitted to the training data, the mean is estimated for the stable (constant) parts of every particular context that is required. After estimating the reference stable values, these values can also be predicted for the unseen samples of the test set. We evaluate the model fit (as described in III-B1) in order to compare the trajectories obtained with predicted stable values.

For the measurements described here, change-descriptor model parts are treated differently. We choose to determine change-descriptor behaviour in terms of temporal information and define two representative parameters to evaluate the consistency: (1) Relative position to ASR boundary and (2) Absolute duration.

During the speech segmentation process, a single ASR boundary for every phone transition is obtained. This boundary has the same location for all 13 cepstra and is useful to provide an initial alignment of similar transition examples. In this way, we compensate for the fact that not all examples are equal in length. Measuring the centre positions (exactly half way between the model boundaries) of the change descriptors and relative to the ASR boundary then provides a good indication to the position where most of the change for a cepstral transition is occurring. The absolute duration is the length of the change descriptor as defined by the model alignments. Both position and duration measurements are given in terms of frame units.

For each cepstral transition class, we estimate:

$$\bar{x}_{cep} = \frac{1}{N} \sum_{n=1}^{N} x \qquad (5)$$

and

$$\sigma_{cep} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (x - \bar{x}_{cep})}, \qquad (6)$$

where x is the measured parameter value, $\bar{x}_{cep}$ the mean and σ_{cep} the standard deviation for N examples of the cepstral transition class. To represent an entire set of cepstal transitions with a single consistency value, we sum the contributions from each class:

$$C_{global} = \frac{1}{T} \sum_{t=1}^{T} \sigma_{cep}, \qquad (7)$$

where σ_{cep} are the standard deviations estimated for N examples cepstral transitions and a total of T classes.

C. *Cepstral model improvements*

In order to gain a better understanding of the trajectory model, and to improve its capabilities, we have refined the basic model along a number of dimensions. These refinements are described below.

1) Connecting model segments: In the standard model, each segment is modelled separately. This means that stable value estimates of two adjacent models will not necessarily be the same, but could exhibit a 'gap'. We extend the piece-wise linear approximation algorithm to model the entire utterance in a single process, eliminating this artificial gap by forcing adjacent stable values to be equal to one another.

2) Predicting stable values with constrained alignments: Trajectory models with fixed reference stable values behave differently with regard to the model alignment algorithm than free trajectory models. An intermediate option would be to first fit free trajectories (to find transitions), then enforce stable values (from predicted reference values). This combined information then constitutes the final trajectory.

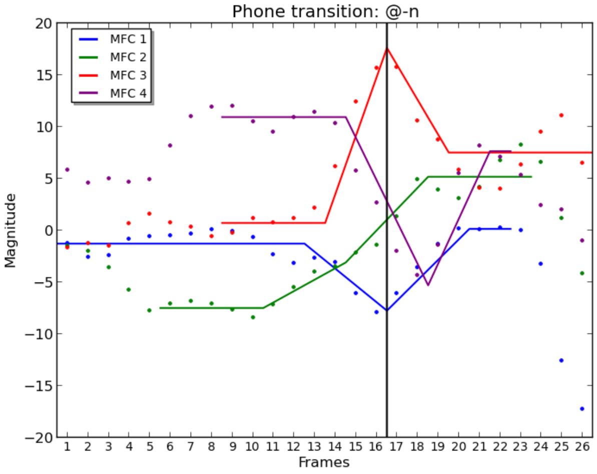

Fig. 2. Piece-wise linear model fit of the first four cepstra of the diphone transition /@-n/ using 4-piece connected models

3) 4-piece models: In Figure 2 another example of the same diphone transition as shown in Figure 1, but in a different context, is shown. While some of the cepstral transitions are seen to be moving (relatively) in similar directions and some start and end positions seem to agree, it is clear that the transition itself behaves rather differently. Instead of single transitional changes, characteristic peaks and troughs are now formed. This behaviour is seen quite frequently for certain transitions and coefficients. It is clear that in such cases, a more elaborate change descriptor could be of value to model the change accurately.

To improve change descriptor representation, we implement a 4-piece symmetrically constrained model. With this configuration, the change descriptors consist of two line pieces, which are kept at equal length. (This requirement drastically reduces the search space to find model alignment, compared to the requirement when any of the four line pieces could be of arbitrary length, and may be more robust in specifically detecting the peaks and troughs). In Figure 2 the linear approximations with 4-pieces are estimated for the first four cepstra. Since this is also a connected model, the stable value parts are now shared with the adjacent transition models (e.g. the stable value of the first cepstral coefficient, MFC 1, is now fitted to frames $1-12$ with this diphone transition example only beginning at frame 9).

IV. Experimental set-up

A. *Overview*

The experiments of this paper are performed on phone transitions that are selected to ensure that data scarcity does not interfere with our investigation. (Although we eventually want to apply our model in limited-data environments, our current goal is to understand its description of speech features in the absence of such a constraint.) In this section, we provide a discussion of the selection process. Each phone transition is selected from a high quality set of speech recordings (of a single speaker) and reviewed acoustically before being included in the final data set. To model any phone transition, specific MFCC features and the appropriate speech segmentation are derived. We also present the specifics of the features used to model the cepstral transitions.

B. *Speech data*

About 6000 short utterances were recorded for the experiments that we conduct. This provides a large corpus of high quality speech of a single male speaker. Only considering a single speaker allows us to focus on contextual effects first, without inter-speaker differences complicating the results. The recordings were made using a list of short Afrikaans prompts (1 to 5 words in length) with balanced phonetic coverage [12]. Additionally, a dynamic programming scoring algorithm was used with initial acoustic models to verify the speaker's pronunciations and obtain a high quality (aligned) set of recordings [13]. The number of utterances that showed perfect alignment was 4974 and had a total duration of about 3 hours.

From this "clean" data set, training and test data sets were selected. All diphone transitions that occur 30 or more times in the clean data set were retained, and greedy selection was used to select test utterances until the test set contained at least 3 examples of each of these diphones. The remaining clean utterances formed the training data set. After performing these steps the total number of utterances that were selected totalled 902 and 4072 for the test and training data sets, respectively.

C. *Segmentation*

Accurate identification of phone transition boundaries is very important, since our modelling approach relies on these boundaries. We use a standard HMM-based ASR system trained on all 4974 clean recordings to automatically align the speech data. A context-dependent cross-word phone recogniser with tied triphone models is employed; 39 MFCC features are used, which include the first 13 and their first and second order derivatives. These features are computed with a window size of 25ms and a frame rate of 10ms. Semi-tied transforms are applied. Each triphone model has 3 emitting states with 7 Gaussian mixtures per state and a diagonal covariance matrix. Verifying phone recognition accuracy on the test set, using a flat-phone grammar yields a value of 92.71%.

Triphone model alignments are obtained using a forced alignment on all the data and the model alignment labels are then converted to the base label sequence (the actual phonemes observed in the training data).

D. Features for transition modelling

After transition boundaries have been obtained, we extract 13 MFCCs features for transition modelling. For these features, while we still use a window size of 25ms, the frame rate is adjusted to 5ms. This provides us with better time resolution. Only the raw MFCC coefficients are used and not any of the derivatives. Finally, for every utterance, each of the MFCC vectors is associated with the phone-boundary alignments from above, which provides contextual labelling at the triphone level.

E. Selection of transition examples

Transitions	**All data**	**Train**	**Test**
Total number	783	769	678
$>$ 30 examples	470	436	173
Final selection	331	331	331

TABLE I
Number of unique diphone transition labels in data sets for various selection stages.

Given the test and training data sets, a further selection process was used to select the data for our experiments. In Table I, the total number of unique transition labels is given to show that for a large number of labels (470) we have more than 30 examples. (For the transition model analysis all transitions with fewer examples are ignored.) After excluding transitions including the silence label, we perform a final (per example) selection. A particular transition example is only allowed if the duration (in frames) is no more than a single standard deviation from the mean. The result of this selection provides us with the 331 most frequent transition labels and transition examples that have low speech rate variability.

V. Experiments and Results

We compare various trajectory tracking techniques, reporting on the results obtained with each of the possible improvements described in Section III-C. For all of the model options, the MSE_{global} values are calculated to measure overall effectiveness. In the case of connected models, we always convert to a valid segmented representation, which ensures that direct comparison of the phone transitions on a per-segment basis is valid. Model options with predicted stable value parts require a train and test data set to assess trajectory tracking. Reference stable values are predicted using the training data and then applied as fixed stable value fits during model estimation on the transitions of the test data set.

To compare change-descriptor behaviour we estimate the global consistency values C_{global} of specific temporal parameters. More detailed comparison can be obtained (on a cepstral level), comparing the standard deviation σ_{cep} for the same cepstral transitions.

A. Connecting segments

As mentioned in Section III-C3, correctly representing the more stable parts of phone transitions given the imposed models requires us to extend the piece-wise linear approximation algorithm. Now, an entire utterance must be represented by a single piece-wise approximation. Finding the utterance-level trajectory model is accomplished in two estimation steps (adding to the definition in Section III-A):

- Locate the model alignments for all of the transitions in the utterance for segmented models.
- Use model alignments to fit all required line pieces of the utterance. On a per-segment basis (left-to-right), fit the change descriptor and ending stable value (except for the first transition), re-using the last stable value of the previous segment as the first stable value of the current segment.

Additionally, if fixed reference stable values are required, fit the mean of the two reference stable values contributing to a single shared stable value. By sharing the first stable value of the previous segment (transition), the segmented models are connected to form a single trajectory for the whole utterance.

Table II shows all of the estimated MSE_{global} values. Global MSEs are estimated for the phone transitions of different data sets (train and test) and two values are given per measurement: the means and standard deviations (in brackets) of the diphone transition class MSEs, respectively. Only free trajectories are constructed for the training data set. The global MSE for the test set, however, are compared for all options (free or fixed stable value trajectories). To aid the comparison, a ratio is also determined between the global MSE values with every fixed stable value trajectory option and its corresponding free trajectory. Finally, separate model parts can be evaluated, and the global MSE values for only the frames corresponding to the stable value of change-descriptor parts of trajectory models are given.

We observe that the error on the training set is in agreement of the test set results for free trajectories. There is a cost to connecting segments: Overall the error increases (as can be expected for the more constrained model). However, the ratios of error between fixed stable value and free trajectories are similar, and we see that the error increases at least five-fold when predicting stable values (rather than estimating them on each phone occurrence).

As previously observed [11], larger context sizes allow for more specific stable values and improved model fit. Therefore, we test reference stable values of different context sizes (monophones, biphones and triphones) and find that predicted stable value model fits improve up to the triphone context

Model	**Stable reference value**	**Global MSE (train)**	**Global MSE (test)**	**Ratio (with free fit)**	**Global MSE (stable values)**	**Global MSE (change)**
3-piece segmented	Monophone		24.553 (6.338)	6.803	27.636 (6.662)	8.861 (5.165)
	Biphone		19.228 (4.223)	5.328	21.908 (4.447)	6.756 (2.979)
	Triphone		19.118 (3.574)	5.297	21.961 (4.105)	6.974 (2.141)
	No ref	**3.604 (1.265)**	**3.609 (1.306)**		**4.047 (1.448)**	**1.821 (0.794)**
3-piece connected	Monophone		47.659 (8.232)	6.196	54.594 (10.458)	18.354 (5.893)
	Biphone		43.595 (7.319)	5.668	50.038 (9.503)	17.339 (4.226)
	Triphone		42.157 (7.064)	5.481	48.826 (9.219)	17.038 (3.512)
	No ref	**7.710 (2.335)**	**7.692 (2.433)**		**8.299 (2.667)**	**5.415 (1.837)**
4-piece connected	Monophone		22.788 (3.614)	5.450	38.089 (6.691)	13.482 (2.362)
	Biphone		21.497 (3.410)	5.142	35.968 (6.109)	12.461 (2.107)
	Triphone		21.398 (3.453)	5.118	36.246 (5.978)	12.303 (2.220)
	No ref	**4.211 (1.243)**	**4.181 (1.242)**		**4.959 (1.481)**	**3.427 (1.106)**

TABLE II
Overall MSE_{global} measurements for train and test data trajectories, including options with predicted stable values.

level. Finally, consistency measures of the mean position of the change descriptor and the mean duration of the change descriptor show similar distributions for 3-piece segmented and connected models.

B. Aligned transitions

Once stable values have been estimated, the timing of the change descriptor of a specific transition is determined by finding the best fit from one stable value to another. This may not produce optimal change descriptor alignment, especially if a specific stable value does not suit a specific sample of a transition well. Change descriptor alignments can be constrained to the free trajectory alignments for better change detection. The fixed stable values can then still be applied without allowing the model to find a further optimal fit for chosen parameters.

In Table III, a global free trajectory baseline consistency (C_{global}) of the change descriptor centre position is estimated. Since trajectories with fixed stable values only exist for the test data, all comparisons are made for the transitions of the test set. We find that the measured change descriptor position is less consistent for trajectory models with fixed stable values (free trajectory models show most consistent change descriptor positions in all cases).

The more consistent change descriptor positions of the free trajectory models motivate further investigation of free trajectory alignments. To better understand the relationship between reference stable values, free trajectory alignments and model fit, we also determine the MSE parameters when constraining fixed stable value trajectory models to have free trajectory alignments. Similarly to the values in Table II, Table IV shows the MSE_{global} values, now with free trajectory alignments. As expected, the overall MSE measurements show increased error for constrained alignments. Since 3-piece model change-descriptors are so dependent on stable values we find substantial error increases when comparing the values of Table II.

Model	**Stable reference value**	**Position (centre)**
3-piece segmented	No ref	2.847 (1.215)
	Biphone	4.095 (1.753)
	Triphone	4.107 (1.753)
3-piece connected	No ref	2.848 (1.215)
	Biphone	3.924 (1.708)
	Triphone	4.024 (1.721)
4-piece connected	No ref	2.167 (0.751)
	Biphone	2.289 (0.832)
	Triphone	2.281 (0.827)

TABLE III
Overall consistency C_{global} measurement of change descriptor position on test set

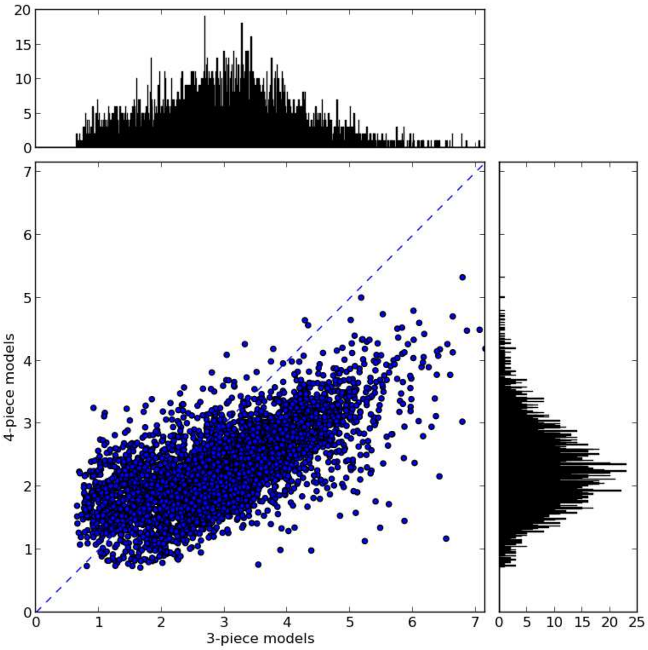

Fig. 3. Comparing consistency σ_{cep} of change descriptor position on a per cepstrum basis

C. 4-piece segments

For all the previous trajectory options, a change descriptor consisted of a single straight line, connected to the start and end points of the stable values (at the model alignments). With 4-piece models, the complexity of the change descriptor is

Model	Stable reference value	Global MSE	Global MSE (stable values)	Global MSE (change)
3-piece connected	Monophone	50.245 (8.563)	54.156 (9.826)	36.059 (7.502)
	Biphone	45.429 (7.315)	49.102 (8.512)	32.014 (5.904)
	Triphone	44.521 (7.057)	48.169 (8.161)	31.199 (5.870)
4-piece connected	Monophone	32.890 (5.485)	50.626 (8.655)	15.527 (2.696)
	Biphone	29.856 (4.771)	45.772 (7.339)	14.268 (2.409)
	Triphone	29.373 (4.888)	45.001 (7.151)	14.002 (2.521)

TABLE IV

Overall MSE_{global} measurement on test set, when applying fixed stable values and constrained alignments

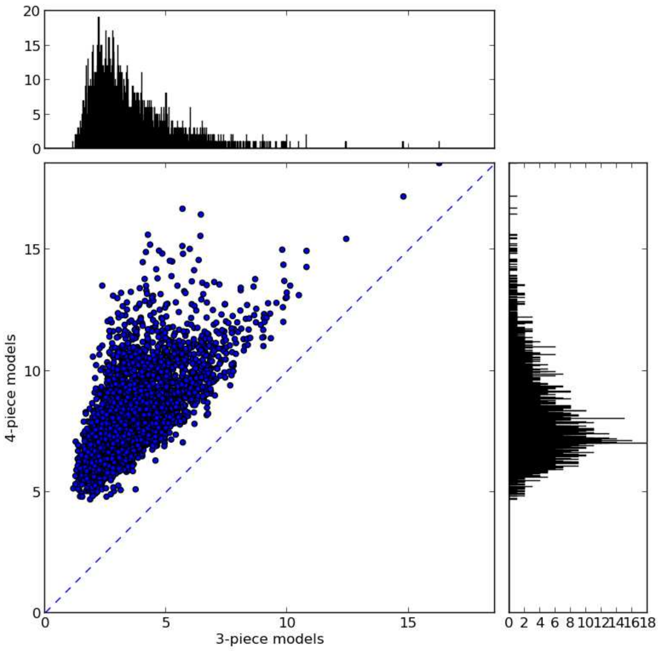

Fig. 4. Comparing mean duration $\bar{x}_{cep}$ of the change descriptors on a per cepstrum basis

increased to include two straight lines. This allows the change descriptor to have a freely varying centre point (connecting the two change descriptor line pieces). As a final constraint, the change descriptor must be symmetrical along the time axis (two lines of equal duration) during model alignment. Final model fits (when connecting segments) of the utterance may however find “shared” stable values different to the ones used during model alignment, leading to non-symmetric change descriptors.

Comparing the overall consistency C_{global} of the change descriptor position with that of the 3-piece models shows the 4-piece models to be the most consistent choice for free trajectory models (Table III). Furthermore, rather than becoming much less consistent when trajectories with fixed stable values are used, 4-piece models show comparable consistency for fixed stable value trajectories.

Figure 3 provides a more detailed comparison. Measuring the centre position of the change descriptors relative to the ASR boundary for each cepstral transition example, and computing the standard deviation σ_{cep} allows transition comparison on a per cepstrum basis. The scatter plot therefore depicts these values for 3 and 4-piece models and the same transition examples. We find that most of the cepstral transitions have larger standard deviations when 3-piece models are used. According to the histogram frequencies and the placement of cepstral transition measurements, only a relatively small number of cepstral transitions have smaller standard deviation for 3-piece models. Generally 4-piece models also tend to have lower standard deviation for most of these cepstral transitions.

Model	Duration (absolute)
3-piece segmented	2.710 (0.858)
3-piece connected	2.583 (0.876)
4-piece connected	2.842 (0.930)

TABLE V

Overall consistency C_{global} measurement of change descriptor durations on all data

To understand more about the differences between the 3-piece and 4-piece change descriptors, we also compare their absolute durations (length in frames). Figure 4 shows the mean duration in frames compared for every cepstral transition class between 3-piece and 4-piece models. It is clear that for all cepstral transition classes, the mean durations of the change descriptors are longer for 4-piece models compared to 3-piece models and the same class.

Confirming the overall variability C_{global} on the mean duration (free trajectories), we find that connecting segments for 3-piece models seems to provide the most consistent mean change descriptor durations in general (Table V). Although 4-piece models with longer change descriptors are less consistent, this value is still very comparable to the 3-piece model case.

Finally, the overall MSE_{global} values in Table II confirm that the additional freedom of the 4-piece model reduces overall error by considerable amounts; this is true for both the model fit of change descriptor and stable value parts, as well as trajectories with predicted stable values compared to 3-piece models of similar configuration. The ratio of the MSE for trajectories with predicted stable values and free trajectories is also seen to improve for 4-piece models.

Additional insight regarding the predictability of reference stable values may be achieved by exchanging (“swapping”) the matching reference values between 3 and 4-piece models. Table VI shows the MSE_{global} values for the different context

Model	Stable reference value	Global MSE	Global MSE (stable values)	Global MSE (change)
3-piece connected	Monophone	47.860 (8.270)	54.505 (10.367)	18.474 (6.062)
	Biphone	43.701 (7.381)	49.706 (9.341)	17.306 (4.309)
	Triphone	42.638 (7.440)	48.940 (9.445)	17.159 (3.738)
4-piece connected	Monophone	23.095 (3.708)	38.847 (6.920)	13.518 (2.393)
	Biphone	22.070 (3.612)	37.100 (6.586)	12.665 (2.184)
	Triphone	21.906 (3.597)	37.245 (6.375)	12.465 (2.271)

TABLE VI

Overall MSE_{global} on test set for “swapping” reference stable values between 3 and 4-piece models

sizes. Improved model fit of stable regions for 3-piece models (using the 4-piece predicted stable values) are obtained, in all cases except the triphone case. Similarly, the 4-piece model model fit for these regions degrades in all cases. Overall model tracking degrades slightly in all cases.

VI. Conclusion

With this work we improve upon the piece-wise linear model approximation of cepstral transitions. This is accomplished by the introduction of new approximation options (connecting segments, constraining model alignments and more complex change descriptors). Trajectory model tracking is analysed in more detail and for separate model parts (change descriptors and stable values). We find that connecting segments, to form a single linear approximation for the entire utterance, proves to be successful and leads to similar distributions for the change descriptors. Although we do obtain similar context dependent improvements to [11] for predicting stable values, these predictors are confirmed not to be very accurate representations of the actual magnitudes for frames of the stable regions of individual transition examples.

Our analysis of change descriptor behaviour shows free trajectories to be the most consistent at detecting the relative position of change. Change descriptor behaviour is tightly coupled to the chosen stable values for 3-piece models and are therefore strongly affected, introducing large error, for aligned model fits. In contrast, the extra degree of freedom for the change descriptor of the 4-piece model is seen to be much less dependent on the stable value parts, resulting in comparatively consistent positions of detected changes. Further examination of the change descriptors shows the 4-piece approximation to model much longer changes in general, which agrees with plots of cepstral transitions where characteristic (longer) double transition behaviour can frequently be observed near the ASR boundary for some cepstra and phone transition labels. This also implies that fewer frames are assigned to stable regions.

In spite of these factors, the error in the stable values of the 4-piece approximation increases substantially for constrained alignments and is fairly similar to the stable value error for the 3-piece model case. Swapping the predicted stable values between 3 and 4-piece models also generate similar error for these parts, with slight improvement when using the 4-piece predictors. The exact reason why these regions show so much intra-segmental variability is not yet well understood and additional investigation may prove valuable.

References

[1] V. Digalakis, “Segment-based stochastic models of spectral dynamics for continuous speech recognition,” Ph.D. dissertation, Boston University, 1992.

[2] W. Holmes and M. J. Russell, “Probabilistic-trajectory segmental HMMs,” *Computer Speech and Language*, vol. 13, no. 1, pp. 3–37, January 1999.

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

[4] 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.

[5] H. V. hamme, “An on-line NMF model for temporal pattern learning. theory and application to automatic speech recognition,” in *Proc. International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2012, pp. 306–313.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] J. Badenhorst, M. Davel, and E. Barnard, “Analysing co-articulation using frame-based feature trajectories,” in *Proc. PRASA*, November 2010, pp. 13–18.

[11] J. Badenhorst, M. Davel, and E. Barnard, “Trajectory behaviour at different phonemic context sizes,” in *Proc. PRASA*, November 2011, pp. 1–6.

[12] N. 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] M. Davel, C. van Heerden, N. Kleynhans, and E. Barnard, “Efficient harvesting of internet audio for resource-scarce ASR,” in *Proc. Interspeech*, August 2011, pp. 3153–3156.