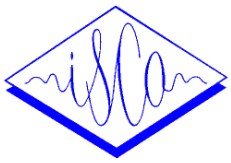

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

STATISTICAL TRAJECTORY MODELS FOR PHONETIC RECOGNITION [1]

William D. Goldenthal and James R. Glass

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
email: thal & jrg @goldilocks.lcs.mit.edu

ABSTRACT

In this work, we apply statistical trajectory models (STM's) to the task of phonetic recognition. STM's attempt to capture the dynamic characteristics and statistical dependencies of acoustic attributes in a segment–based framework. The approach is based on the creation of a *track*, $\vec{T}_\alpha$, for each phonetic unit α. The track serves as a model of the dynamic trajectories of the acoustic attributes over the segment. The statistical framework for scoring incorporates the auto- and cross-correlation properties of the track error over time, within a segment. This paper presents the results of a series of phonetic recognition experiments using the TIMIT acoustic–phonetic corpus [1]. Using the NIST train and core test sets we obtained context–independent and context-dependent recognition accuracies of 64.0% and 69.0% respectively.

INTRODUCTION

The objective of our work is to develop segment–based models which capture the time–varying characteristics of the acoustic attributes used to represent the speech signal. Segment–based approaches hypothesize start and end times for each segment during the matching process. While assuming independence between adjacent segments, these formulations offer the possibility of directly modelling the within–segment correlations. For each hypothesized N frame speech segment, $S = \{\vec{s}_1, \ldots, \vec{s}_N\}$, an N frame synthetic segment G_α is generated for each phonetic model α. Scoring, performed at the segment level, is then based on the error sequence, E, determined by comparing each G_α to S.

In the next section, we describe our method for generating a synthetic segment and for creating a statistical model. We then present our attempts to model co–articulatory effects by *merging* tracks, and also by modelling phonetic transitions. This is followed by a description of context–independent and context–dependent phonetic recognition experiments using the TIMIT acoustic–phonetic corpus [1]. Finally, the results are discussed and summarized.

MODELLING SPECTRAL DYNAMICS

In this work, dynamics are modelled by creating a phone dependent *track* which can be defined as a trajectory, or temporal evolution of the acoustic attributes over a segment. A track consists of a sequence of M state vectors $T = \{\vec{t}_1, \ldots, \vec{t}_M\}$ which are used as the basis for generating a synthetic segment

$$G = f(T, N) = \{\vec{g}_1, \ldots, \vec{g}_N\} \quad (1)$$

for any number of frames N, where $f()$ is a generation function.

To classify an N frame speech segment, S, we generate synthetic segments, G for each phonetic model α. The match between S and G is made by computing the error

$$E = S - G = \{\vec{e}_1, \ldots, \vec{e}_N\} \quad (2)$$

where

$$\vec{e}_i = \vec{s}_i - \vec{g}_i \quad (3)$$

The likelihood of each phonetic model, α, is based on the error probability $P(E|\alpha)$. The statistical framework for scoring incorporates the auto- and cross-correlation properties of the track error over time, within a segment. Hence, statistical dependencies are incorporated into the model.

There exist other approaches in the literature which attempt to explicitly model the segment level dynamics [2, 3]. These approaches are based on parametric models of the dynamics. In this work we do not parameterize the spectral trajectories. Instead, we create tracks from the data by mapping it to a sequence of M states for each phone. When all the tokens in the training set for a particular phone have been mapped, the phone dependent track is calculated from the ML estimate of each state. The objective is to capture the non–linear nature of the dynamics in a simple way, without explicitly assuming anything a–priori about the dynamic properties of the signal.

The choice of the generation function, which accounts for the durational variability of a phone, is described in more detail elsewhere [4]. For this work, the generation function is a linearly interpolated mapping of a token's frames to the states of the track. The initial and final frames of the token are always aligned with the initial and final states of the track.

ERROR MODELLING

The objective of the error model is to take advantage of information residing in the error correlations both over time and between attributes. In [2] the errors are assumed to be independent within a segment, while in [5, 6] a joint-Gaussian distribution was used to model the segment.

[1] This research was supported by ARPA under Contract N00014-89-J-1332 monitored through the Office of Naval Research. W.D. Goldenthal receives support from C.S. Draper Laboratory.

Two key difficulties exist in modelling the error with a Gaussian distribution. The first problem is due to the fact that the error sequence varies in duration. The second problem arises when the dimension of the Gaussian distribution becomes large, and the estimate of the covariance matrix parameters become suspect.

The method which we found to provide a good trade-off is to allow the error vectors to be of varying frame length and to normalize the frame length by averaging the vectors over each of Q pieces [4, 7]. For example, for a ten state track with Q equal to three, that part of the error which resulted from comparing the token to the first third of the track (i.e. the first three and a third "states") would be averaged, and so on for each of the other two thirds. This technique has the advantages of reducing the dimensionality of the Gaussian distribution while utilizing all of the data. The disadvantage is the loss of the fine detail of the error correlations.

Therefore, our model of the error is the maximum likelihood estimate of the mean (which is not zero, due to the averaging into Q pieces) and full covariance matrix for each phone. For P acoustic attributes, the result is a joint-Gaussian density of dimension PQ. Note that the dimension of the model is independent of the number of states, M, used to characterize the track.

MODELLING COARTICULATION WITH MERGED TRACKS

Although contextual influences can extend across several phonetic boundaries, it is most often the case that the effect of the phone in the left context position is primarily seen near the left boundary of a segment, and the effect of a right context phone is primarily seen near the right boundary. Therefore, as a first approximation, it will be assumed that the contextual effects of the phonetic environment can be modelled as influencing primarily the adjacent region of a segment. This simplification should enable us to capture the dominant contextual effects in a novel and efficient way. A method which will take advantage of the dynamic tracks is to *independently* account for the left and right contexts by creating biphone tracks, and then combine them to create triphone tracks as they are needed. Tracks can be estimated and stored for the left and right contexts separately and then *merged* when a synthetic segment is generated to create a triphone based synthetic segment.

This implementation dramatically reduces the magnitude of both the coverage problem and the sparse data problem. For a system with $N = 58$ phonetic models, the maximum number of required context–dependent tracks is reduced from $N^3 = 195{,}112$, to $2*N^2 = 6{,}728$, not accounting for the large number of transitions that never occur in English. The factor of two occurs because for a given transition we get two possible tracks (e.g. for [rɑ] the [r] data is used for modelling a right context track for [r], and the [ɑ] data is used for modelling a left context track for [ɑ]).

Additionally, since the error modelling techniques are independent of the track, the number of statistical error models (which require the majority of parameters) is a design parameter, since the errors can be pooled over all contexts. Therefore only the track will be context–dependent. This pooling of the error matrices will dramatically reduce the impact of any sparse data difficulties with respect to the estimation of the statistical parameters.

Hence, the main ideas are first, to generate robust biphone tracks, second to merge these tracks to generate triphone synthetic segments, and third to use the errors generated from these triphones to estimate the error covariance parameters. Finally, since the left and right contexts are utilized independently, triphone tracks can be created as needed during test. That is, triphone contexts never seen during training can be *synthesized* from the left and right biphone tracks. This presents a possible method of greatly increasing the coverage provided by the training set.

The procedure for generating the context–dependent biphone tracks consists of first constructing a track for all biphone combinations in the training set. These tracks are then clustered bottom–up, using a track distance metric (TDM). This metric is defined as follows. Let P represent the number of acoustic attributes, M represent the number of states, and N represent the count for each state in a track. Then, given the merger weight for state i, w_i, the distance between two tracks for phone α in the two contexts β and γ, is TDM(T_β, T_γ):

$$TDM(T_\beta, T_\gamma) = \sum_{i=1}^{M} w_i \left[\frac{N_{\beta_i} * N_{\gamma_i}}{N_{\beta_i} + N_{\gamma_i}}\right] \sum_{j=1}^{P} \left[\frac{(T_{\beta_{ij}} - T_{\gamma_{ij}})^2}{\sigma^2_{\alpha j}}\right] \quad (4)$$

where $\sigma^2_{\alpha j}$ is the phone dependent variance of the j^{th} dimension.

MODELLING PHONETIC TRANSITIONS

Another method of using dynamic tracks to enhance system performance is to examine the acoustic information that spans adjacent segments. The idea is to make tracks of the phonetic transitions themselves. This lends itself well to the overall approach, since the transition regions are highly dynamic because the articulators are generally in motion during this interval. During recognition, the transition model scores augment the segment scores to provide contextual information.

The main difficulty which needs to be overcome is the very large number of phonetic transitions which occur. Sparse data considerations limit the number of models which can be created. However, many transitions are very similar. While it may be impractical to capture all of the transitions, it may be possible to create a significantly large subset of transition models by clustering the transitions, and then pooling the data according to the clusters which are formed.

Other approaches have attempted to utilize a method of explicitly scoring the phonetic boundaries [8, 9, 10, 11]. These approaches tend to use broad phonetic classes, often based on place–of–articulation, to create a manageable number of models. The idea to be explored in this paper is to use the TDM to cluster together transition tracks to arrive at a group of transitions which are representative of the major classes which such transitions fall into. These major transition classes would essentially be *canonical* transitions. However, rather than use a predetermined set of broad linguistic categories, bottom–up clustering will again be employed. This will allow a large number of unsupervised data–driven transition models to be created.

The transitions help in two ways. First, the transition

scores will be incorporated into the overall scoring framework to help determine the phonetic identity of the two phones involved. Secondly, they can be examined to determine likely segment boundaries within an utterance. This reduces the possible search space when we do phonetic recognition, particularly since the transition likelihoods provide an idea of which phones are involved in the transition.

EXPERIMENTAL RESULTS

Corpus and Representation

The recognition experiments are based on the TIMIT acoustic–phonetic speech corpus [1]. Our training and test sets were chosen to facilitate comparison of the results with other work reported in the literature. The primary sets were the NIST designated "core" test set consisting of 24 speakers, and the NIST designated training set, consisting of 462 speakers. These sets were constructed such that the sentences are completely disjoint from each other. This lack of overlap between training and test set utterances precludes the grammar model from containing a favorable statistical bias towards the overlapping sentences in the test set.

All of the experiments used Mel-frequency cepstral coefficients (MFCC's) to represent the speech signal [12]. The MFCC's were used to facilitate comparison to previously published results [2, 14]. Our experiments also made use of ΔMFCC's which were computed at the beginning and end of the phonetic segment [5]. The context–independent experiments made use of 15 MFCC's while the context–dependent experiments used only 13. Segment duration was also included as a parameter. The recognition experiments were conducted using a segment–based Viterbi search, with phonetic boundaries hypothesized every 10 ms.

The gender of the speaker was assumed to be known. However, we do not believe this is a strong assumption. Studies by Lamel and Gauvain conducted on the TIMIT corpus showed accurate gender identification performance of better than 97% after 0.4s of speech (about four phones), which improved to 99% after 2.0s [13]. When errors were made, the results were better for the cross–gender models.

Track Creation

A gender specific phonetic model was created for each of the phone labels used in the TIMIT data base, with some collapsing of silence labels and syllabic nasals. Each of the tracks used in the models contained 10 states. The statistical components used a value for Q of three resulting in a 76 and 66 dimensional distribution for context–independent and context–dependent models respectively.

Transition tracks consisting of 21 states were computed for each of the 1,275 distinct transitions found in the training set. The tracks were then clustered in an unsupervised manner, with the same bottom–up algorithm used in the creation of the context–dependent tracks. To allow for a larger number of canonical transitions, the ΔMFCC's were not used as statistical features. Hence, the total dimension of the Gaussian distribution is 45 (no duration component), resulting in 1,013 independent parameters in each covariance matrix. This allowed for the creation of 200 canonical transition models for the NIST training set.

STM Results on the NIST "Core" Test Set		
Condition	Accuracy	
	w/o transitions	with transitions
Context–Independent Bigram	61.9%	64.0%
Context–Dependent Constrained Trigram	66.4%	69.0%

Table 1: Context–independent and context–dependent recognition results for the NIST Core Test Set using 39 phone classes. The CI results use a bigram grammar and the CD results use a "constrained trigram" grammar (see text for details).

Context–Independent Phonetic Recognition

The context–independent experiments provide a baseline standard upon which subsequent improvements can be measured. To refine the system design parameters, models were trained on the NIST training set and initial evaluation was performed using a development set consisting of 50 speakers not used in the NIST training and test sets. The grammar was a phone bigram, based on the same 58 phonetic units for which acoustic models were constructed.

The results of the context–independent experiments, both with and without the use of the phonetic transition models, are presented in Table 1. The results are based on the 39 classes defined by Lee and Hon [15]. To utilize the transition models in a context–independent manner, the log likelihoods for all the transitions from the phone being scored were exponentiated and summed. This results in a transition likelihood which is conditioned only on the hypothesized phone.

Context–Dependent Phonetic Recognition

The models for this set of experiments were trained on the NIST training set. A total of 2,492 gender dependent tracks were generated via bottom–up clustering, using the TDM. The errors for each phone were pooled into a single covariance matrix for each of the gender models. Since the context–dependent experiments impose a significantly larger computational burden than the context–independent experiments, the system design parameters were tuned using a smaller development set consisting of only 16 speakers. When statistical measures (e.g. mean track distortion) were used to choose design parameters (e.g. the mean track distortion is used to determine the threshold for clustering), the full 50 speaker development set was employed.

The Viterbi search was conducted by hypothesizing alternative right contexts for each phonetic model at each point in time (i.e. at 10 ms intervals). The left context and boundary for each model were constrained to be those used in the best path achieved up to the current time. Hence, the search was over all possible bigrams. However, since the left context was "known" for each hypothesized segment, the acoustic models utilized a full triphone track, created by merging the "known" left context biphone track with each allowable right biphone track. Additionally, a trigram phonetic grammar was also employed. Hence, this implementation strategy allowed the use of triphone and trigram information while maintaining a bigram search. We call this a *constrained trigram* search.

The results of the context–dependent experiments, both with and without the use of the phonetic transition models,

Comparison of STM to other Work			
Work	Accuracy	Relevant STM Accuracy	
		w/o transitions	with transitions
KFL (KFL Test Set)	53.3%	63.6%	65.3%
Dynamic Systems (BU Test Set)	63%	66.2%	68.4%
SSM (BU Test Set)	66.7%	67.2%	69.5%

Table 2: Comparison of context–independent recognition results achieved using statistical trajectory models to other work reported in the literature.

are presented in Table 1. In these experiments, the transition likelihoods were used to help determine the right context. Two sets of 200 transition models were constructed. One set was composed only of data from the male training speakers and was used on the male test utterances. The second set used all of the utterances in the training set and was used for the female test speakers.

DISCUSSION

To provide an additional basis of comparison between STM and other approaches, models were also constructed for use with other test sets used in the research community. Early successes in phonetic recognition using HMM's were achieved by Lee and Hon [15]. We used their test set of 20 speakers, however our training set was somewhat larger (the remaining 610 speakers), due to the availability of additional TIMIT data. Segment–based phonetic recognition results have also been reported for the Stochastic Segment Model [6], and the Dynamic Systems model [2]. To compare with their results we tested on their set of 16 male speakers (from the western dialect region), training on the remaining 426 male speakers in the TIMIT corpus.

The results of these context–independent experiments are shown in Table 2. The accuracies in Table 2 are computed using the same criteria for allowable confusions which is used in the work the result is being compared with. Without transition models, the STM results are closest to those of the stochastic segment model (SSM). Both of these approaches utilize some degree of modelling of temporal correlation information.

It is interesting to observe that our results were considerably better on these smaller test sets than the NIST core test set. We believe this reflects both the fact that there is no overlap of sentences in the training and test sets, and that the NIST test set is carefully balanced across all eight dialect regions of the TIMIT corpus.

The STM results in Tables 1 and 2 show an improvement of around 2% in accuracy due to the transition models. In Table 1 the results due to contextual information in the tracks were 4.5% to 5.0% greater than those achieved using context–independent models. The best result of 69.0% is very close to that achieved by Lamel and Gauvain of 69.1% [14]. Their result represents the state–of–the–art in HMM phonetic recognition, which is currently the dominant speech recognition technology. Robinson, using artificial neural networks, has achieved an accuracy of 73.9% [16].

SUMMARY

The investigation reported here represents a preliminary attempt to utilize statistical trajectory models (STM's) for the task of phonetic recognition. Merging of biphone tracks appears to provide a viable means of capturing contextual factors while maintaining robust statistical models. The STM approach is well suited for capturing the high degree of dynamic activity in the phonetic transitions. The transition models clearly provided important additional information which is useful to the recognition process, and significantly boosted system performance. We believe additional refinements of the various STM components will result in additional performance gains.

REFERENCES

[1] L. Lamel, R. Kassel, S. Seneff, "Speech database development: design and analysis of the acoustic-phonetic corpus," *Proc. DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, 100–109, February, 1986.

[2] V. Digilakis, "Segment-based stochastic models of spectral dynamics for continuous speech recognition." Ph.D. Thesis, Boston University, 1992.

[3] H. Gish and K. Ng, "A segmental speech model with applications to word spotting," *Proc. ICASSP 93*, pp. 447–450, Minneapolis, MN, April, 1993.

[4] W. Goldenthal, "Statistical trajectory models for phonetic recognition," Ph.D. Thesis, Massachusetts Institute of Technology, expected September 1994.

[5] H.C. Leung, B. Chigier, J.R. Glass., "A comparative study of signal representations and classification techniques for speech recognition," *Proc. ICASSP 93*, 680–683, Minneapolis, MN, April, 1993.

[6] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition," *IEEE Trans. ASSP*, Vol. 4, No. 12, 1857–1869, December, 1989.

[7] W. Goldenthal and J. Glass, "Modelling spectral dynamics for vowel classification," *Proc. Eurospeech 93*, pp. 289–292 Berlin, Germany, September, 1993.

[8] P. Marteau, G. Bailly, and M. Janot–Giorgetti, "Stochastic model of diphone–like segments based on trajectory concepts," *Proc. ICASSP 88*, pp. 615–618, Tokyo, Japan, April, 1988.

[9] H. Leung. I. Hetherington, and V. Zue, "Speech recognition using stochastic explicit–segment modeling," *Proc. Eurospeech 91*, pp. 931–934, Genova, Italy, September, 1991.

[10] O. Kimball, M. Ostendorf, R. Rohlicek, "Recognition using classification and segmentation scoring," *Proc. DARPA Speech and Natural Language Workshop*, pp. 197–201, Harriman, N.Y., February, 1992.

[11] M. Phillips and J. Glass, "Phonetic transition modelling for continuous speech recognition," *J. Acoust. Soc. Amer.*, Vol. 95, No. 5, pp. 2877, June, 1994.

[12] P. Mermelstein and S. Davis, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. ASSP*, Vol. 23, No. 1, 67–72, February, 1975.

[13] L. Lamel and J. L. Gauvain, "Identifying non–linguistic speech features," *Proc. Eurospeech 93*, pp. 23–30 Berlin, Germany, September, 1993.

[14] L. Lamel and J. L. Gauvain, "High performance speaker–independent phone recognition using CDHMM," *Proc. Eurospeech 93*, pp. 121–124, Berlin, Germany, September, 1993.

[15] K. F. Lee and H. W. Hon, "Speaker–independent phone recognition using hidden Markov models," *IEEE Trans. ASSP*, Vol. 37, No. 11, pp. 1641–1648, November, 1989.

[16] T. Robinson, "Several improvements to a recurrent error propagation phone recognition system," *Technical Report CUED/TINFENG/TR.82*, Cambridge University Engineering Dept., September, 1991.