MAXIMIZING THE CONTINUITY IN SEGMENTATION – A NEW APPROACH TO MODEL, SEGMENT AND RECOGNIZE SPEECH

Ji Ming

Institute of Electronics, Communications and Information Technology
Queen's University Belfast, Queen's Island, Belfast BT3 9DT, UK

ABSTRACT

This paper presents a new approach to speech modeling and recognition. The new approach consists of a statistical model to represent up to sentence-long temporal dynamics in the training data, and an algorithm to identify the matching segments with maximum continuities between the training and testing sentences. Recognition is performed by combining the *longest* matching segments found from the training sentences. Because of their richer and more distinct temporal dynamics, longer speech segments as whole units can be recognized with lower error rates than shorter speech segments. Therefore basing recognition on the longest matching segments optimizes the discrimination and hence recognition of speech. The new approach has been evaluated on the TIMIT database for identifying matching speech segments. The results obtained are encouraging given the very low parametric complexity of the new model.

Index Terms— temporal dynamics, speech modeling, speech segmentation, speech recognition

1. INTRODUCTION

A speech signal has two distinct features: its temporal dynamics, subject to acoustic, lexical and language constraints, and its speaker characteristics. These two features separate a speech sentence from non-speech noise, and from other speakers' sentences. Most conventional speech recognition systems use context-dependent phones as the primary units to model and recognize speech. This has proven to be very effective, in terms of the richness of the acoustic-phonetic training data in many reasonably-sized databases. However, phonetic models are only capable of representing the neighboring phonetic contexts. They lack the ability to represent long-range temporal dynamics, which determines how the individual phonetic sounds are dependent on one another to form a realistic speech sentence. Losing this information, and given the short duration of and hence limited discrimination between the phonetic sounds, conventional speech recognizers lack the robustness to noise.

Over the past years, how to effectively model long-range temporal dynamics in speech has been a research focus. Various approaches have been proposed as an alternative to the conventional, phone-based hidden Markov model (HMM) framework. Typically, these new approaches include the segmental approaches, the dynamic system approaches, and the example-based approaches. In segmental approaches (e.g., [1][2]), phones or phonetic sequences are modeled as segments to capture the dependence within these structures. In dynamic system approaches, some form of linear or nonlinear dynamic system is used to represent the underlying dynamics of speech (e.g., [3]–[5]). More recently, templates or example-based approaches have received renewed interest for their capability to represent long-range temporal dynamics in speech (e.g., [6]–[8]). In contrast to other statistical modeling approaches, templates involve fewer assumptions/manipulations on the speech data and, thus, may be capable of more precisely representing the dynamics within speech. However, unlike statistical models, templates lack smoothness (and hence robustness) in representing short-time speech spectra, which are subjected to random variations.

In this paper, we add a new solution to the problem. We study the *optimal* extraction of long-range temporal dynamics in speech for speech recognition. We optimize the extraction by *maximizing* the size of the primary units to be identified in the testing speech. Larger primary units (e.g., phone sequences, syllables or words) contain richer and more distinct dynamics than individual phones, and thereby they can be recognized with lower error rates. Therefore maximizing the primary units to be identified effectively optimizes the discrimination and hence recognition of speech. In this paper, we describe a system that implements the proposed optimization. The system consists of two parts: (1) a method to model the complete temporal dynamics in each training sentence, such that any segment of any length in the sentence, up to the complete sentence, can be used as a whole unit to identify the corresponding segments/units in the testing speech, and (2) algorithms to identify matching segments with large continuities, and to perform recognition based on the longest matching segments between the training and testing sentences. The difference between our new approach and the conventional template-based approaches will become clear during the course of the description.

2. MODELING LONG-RANGE TEMPORAL DYNAMICS

First, we model the *complete* temporal dynamics in each training sentence, such that any segment of any length in the sentence, up to the complete sentence, can be used as a whole unit to identify the corresponding segments/units in the testing sentence. We use a new example-based approach, as opposed to conventional templates, to build the models. We first train a Gaussian mixture model (GMM) using all the training sentences. Then, based on the GMM, we further build a model for each specific training sentence to represent the full temporal dynamics in that sentence. Denote by G the GMM trained on all the training data for short-time speech spectra x:

$$G = \{g(x|m), w(m) : m = 1, 2, ..., M\} \quad (1)$$

where $g(x|m)$ is the m'th Gaussian component and $w(m)$ is the corresponding weight. Let $\mathbf{x} = \{x_i : i = 1, 2, ..., I_{\mathbf{x}}\}$ be a training sentence represented by $I_{\mathbf{x}}$ frames and x_i being the frame at time i. We can obtain a new representation for $\mathbf{x}$ by taking each frame from the sentence and finding the Gaussian component in

This work was supported by the UK EPSRC under grant EP/G001960.

G that maximizes the likelihood of the frame. This results in a time sequence of maximum-likelihood Gaussian components as a model for $\mathbf{x}$, alternative to the template. In the model, the individual Gaussian components represent the probability distributions of short-time speech spectra at different times, and the time sequence represents the complete temporal dynamics governing the short-time spectra, from acoustic to lexical and to language to form the specific training sentence. Let $g(x|m_{\mathbf{x},i})$ be the Gaussian component identified from G that maximizes the likelihood of frame x_i in training sentence $\mathbf{x}$. Then we can express the new model – the maximum-likelihood Gaussian sequence – for training sentence $\mathbf{x}$ by using the corresponding index sequence $\mathbf{m}_\mathbf{x}$:

$$\mathbf{m}_\mathbf{x} = \{m_{\mathbf{x},i} : i = 1, 2, ..., I_\mathbf{x}\} \tag{2}$$

It may be noticed that the above model G and $\mathbf{m}_\mathbf{x}$ for training sentence $\mathbf{x}$ is similar to an HMM. Indeed, each $g(x|m_{\mathbf{x},i})$ in the model can be viewed as an emission probability density and the sequence index i can be viewed as an index of the state. With a left-to-right state transition, the model forms a time sequence of Gaussian probability densities that characterizes, statistically, the full temporal-spectral activities in $\mathbf{x}$. Recently, there are studies into example-based approaches to speech recognition (e.g., [6]–[8]). These seek more accurate representations of long-range temporal dynamics of speech by making fewer assumptions about the speech. The above model (1) and (2) combines statistical and template-based approaches seamlessly in the same framework, representing a balance between a smooth representation for the short-time spectra and a sentence-long representation for the temporal dynamics.

3. IDENTIFYING MATCHING SEGMENTS WITH LARGE CONTINUITIES

In recognition, we compare the testing sentence with each of the training sentence models to identify *all* their matching segments. Then we form a recognition by combining the *longest* matching segments found from the training sentences. Because of their richer and more distinct temporal dynamics, longer speech segments as whole units can be registered more accurately than shorter speech segments. Therefore basing recognition on the longest matching segments increases the chance of correct recognition. In the following, we describe an algorithm for identifying matching segments with large continuities.

Let $\mathbf{y} = \{y_t : t = 1, 2, ..., T\}$ be a testing sentence with T frames, and $y_{\tau:t} = \{y_\epsilon : \epsilon = \tau, \tau+1, ..., t\}$ be a testing segment in $\mathbf{y}$ consisting of consecutive frames from time τ to t. Let $m_{\mathbf{x},u:v} = \{m_{\mathbf{x},i} : i = u, u+1, ..., v\}$ represent a training segment from model $\mathbf{m}_\mathbf{x}$, addressing the Gaussian sequence modeling consecutive frames from u to v in training sentence $\mathbf{x}$. We measure the similarity between the two segments, $y_{\tau:t}$ and $m_{\mathbf{x},u:v}$, by using the posterior probability of the training segment $m_{\mathbf{x},u:v}$ given the testing segment $y_{\tau:t}$. Assuming an equal prior probability for all the training segments, the posterior probability may be expressed as

$$\begin{aligned} P(m_{\mathbf{x},u:v}|y_{\tau:t}) &= \frac{p(y_{\tau:t}|m_{\mathbf{x},u:v})}{p(y_{\tau:t})} \\ &= \frac{p(y_{\tau:t}|m_{\mathbf{x},u:v})}{\sum_{\mathbf{x}'}\sum_{u',v'} p(y_{\tau:t}|m_{\mathbf{x}',u':v'}) + p(y_{\tau:t}|\phi_{\tau:t})} \end{aligned} \tag{3}$$

where $p(y_{\tau:t}|m_{\mathbf{x},u:v})$ is the likelihood function. We can calculate the likelihood function by using the Viterbi algorithm, which will find the most-likely time map between the two segments. In the calculation, we assume that the frames within a segment are independent of one another. As such, $p(y_{\tau:t}|m_{\mathbf{x},u:v})$ can be written as

$$p(y_{\tau:t}|m_{\mathbf{x},u:v}) = \max_{i_\epsilon} \prod_{\epsilon=\tau}^{t} g(y_\epsilon|m_{\mathbf{x},i_\epsilon}) \tag{4}$$

where i_ϵ is the time map function assuming $i_\tau = u$ and $i_t = v$. As with the usual Viterbi algorithm, we allow a testing segment $y_{\tau:t}$ with length $l_{\tau:t} = t - \tau + 1$ to be compared with training segments $m_{\mathbf{x},u:v}$ with variable lengths from $v - u + 1 = l_{\tau:t}/2$ to $2l_{\tau:t}$, to search for the matching training segment.

In the denominator of (3), the first term corresponds to all the training segments, of variable origins and lengths from all the training sentences, that are likely to match the testing segment $y_{\tau:t}$. The second term corresponds to the likelihood that $y_{\tau:t}$, as a whole unit, matches a segment $\phi_{\tau:t}$ that is unseen in the training sentences, assuming an equal prior as $m_{\mathbf{x},u:v}$. This likelihood can be suitably formed on the universal GMM (1). The following shows a model capable of modeling the likelihoods for arbitrary speech segments:

$$p(y_{\tau:t}|\phi_{\tau:t}) \simeq \prod_{\epsilon=\tau}^{t} \sum_{m=1}^{M} w(m) g(y_\epsilon|m) \tag{5}$$

The posterior probability defined in (3) has an important characteristic: it favors the continuity of the matching segments, in terms of giving larger values for longer matching segments compared as whole units. To show this, assuming that $y_{\tau:t}$ and $m_{\mathbf{x},u:v}$ are a pair of matching segments such that likelihoods $p(y_{\tau:t}|m_{\mathbf{x},u:v}) \geq p(y_{\tau:t}|m_{\mathbf{x}',u':v'})$ for any $m_{\mathbf{x}',u':v'} \neq m_{\mathbf{x},u:v}$, and $p(y_{\tau:t}|m_{\mathbf{x},u:v}) \geq p(y_{\tau:t}|\phi_{\tau:t})$. Express $y_{\tau:t}$ as a union of two consecutive subsegments $y_{\tau:\epsilon}$ and the complement $y_{\epsilon+1:t}$, and $m_{\mathbf{x},u:v}$ as a union of the corresponding matching training subsegments $m_{\mathbf{x},u:i_\epsilon}$ (for $y_{\tau:\epsilon}$) and $m_{\mathbf{x},i_{\epsilon+1}:v}$ (for $y_{\epsilon+1:t}$). We can have

$$\begin{aligned} \frac{p(y_{\tau:t}|m_{\mathbf{x},u:v})}{p(y_{\tau:t}|m_{\mathbf{x}',u':v'})} &= \frac{p(y_{\tau:\epsilon}|m_{\mathbf{x},u:i_\epsilon})p(y_{\epsilon+1:t}|m_{\mathbf{x},i_{\epsilon+1}:v})}{p(y_{\tau:\epsilon}|m_{\mathbf{x}',u':i'_\epsilon})p(y_{\epsilon+1:t}|m_{\mathbf{x}',i'_{\epsilon+1}:v'})} \\ &\geq \frac{p(y_{\tau:\epsilon}|m_{\mathbf{x},u:i_\epsilon})}{p(y_{\tau:\epsilon}|m_{\mathbf{x}',u':i'_\epsilon})} \end{aligned} \tag{6}$$

This because $p(y_{\epsilon+1:t}|m_{\mathbf{x},i_{\epsilon+1}:v})/p(y_{\epsilon+1:t}|m_{\mathbf{x}',i'_{\epsilon+1}:v'}) \geq 1$ based on the assumption that $y_{\epsilon+1:t}$ and $m_{\mathbf{x},i_{\epsilon+1}:v}$ match. In a similar way, we can have an inequality concerning the likelihood ratio associated with $\phi_{\tau:t}$:

$$\frac{p(y_{\tau:t}|m_{\mathbf{x},u:v})}{p(y_{\tau:t}|\phi_{\tau:t})} \geq \frac{p(y_{\tau:\epsilon}|m_{\mathbf{x},u:i_\epsilon})}{p(y_{\tau:\epsilon}|\phi_{\tau:\epsilon})} \tag{7}$$

Applying (6) and (7) to (3) we can obtain two inequalities concerning the posterior probability:

$$P(m_{\mathbf{x},u:i_\epsilon}|y_{\tau:\epsilon}) \leq P(m_{\mathbf{x},u:v}|y_{\tau:t}) \tag{8}$$

$$P(m_{\mathbf{x},u:i_\epsilon}|y_{\tau:\epsilon})P(m_{\mathbf{x},i_{\epsilon+1}:v}|y_{\epsilon+1:t}) \leq P(m_{\mathbf{x},u:v}|y_{\tau:t}) \tag{9}$$

Inequalities (8) and (9) indicate that the posterior probability increases with the continuation of the matching segments, and with comparing successive matching segments as whole units rather than as isolated units. Large posterior probabilities, thus, indicate large continuities between the matching training and testing segments.

Given a testing sentence, we will compute the posterior probability $P(m_{\mathbf{x},u:v}|y_{\tau:t})$ for every testing segment $y_{\tau:t}$ against every training segment $m_{\mathbf{x},u:v}$ for every training sentence $\mathbf{x}$. This full search will expose all matching segments between the training and testing sentences of arbitrary length up to the complete sentences. We will form a recognition based on the longest matching training segments, indicated by the maximum posterior probabilities.

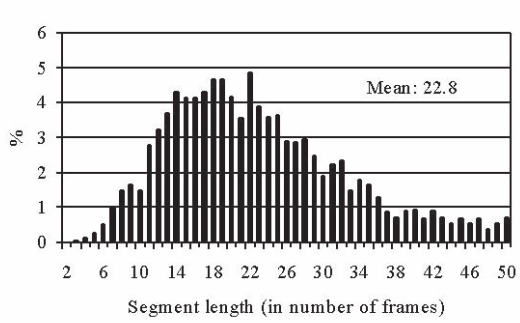

Fig. 1. Histogram of the length of the optimal segments identified on the testing sentences.

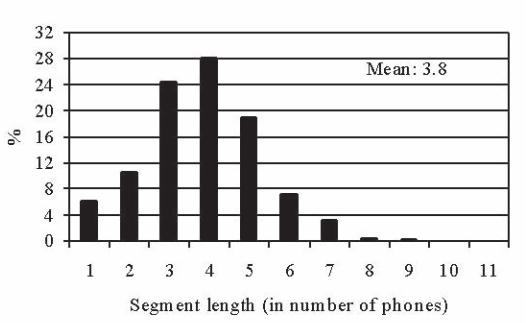

Fig. 2. Histogram of the number of phones spanned by the optimal segments.

4. RECOGNITION BASED ON OPTIMAL SEGMENTS

We use dynamic programming (DP) to combine the training segments with large posterior probabilities (and hence large continuities) into a complete sentence, as a recognition of the testing sentence. We create a posterior probability for the complete testing sentence by concatenating the segmental posteriors $P(m_{\mathbf{x},u:v}|y_{\tau:t})$; DP is used to maximize this sentence posterior by optimizing the starting time τ and ending time t of each segment. This will result in a recognition focusing long matching segments between the training and testing sentences. As these large matching segments are compared as whole units, we expect increased acoustic discrimination.

Let $\delta(t)$ represent a partial logarithmic posterior ending at time t. We have the following recursion for an optimal segmentation for the testing sentence:

$$\delta(t) = \max_{\tau}[\delta(\tau-1) + l_{\tau:t} \max_{\mathbf{x}} \max_{u,v} \ln P(m_{\mathbf{x},u:v}|y_{\tau:t})] \quad (10)$$

The maximization inside the brackets seeks the training segment that best matches the given testing segment $y_{\tau:t}$; the maximization outside the brackets maximizes the continuity of the matching segments over the complete sentence. In (10), $l_{\tau:t}$ is the length of segment $y_{\tau:t}$, introduced to normalize the sentence scores across different segmentations. We form the recognition by concatenating the optimal training segments $m_{\mathbf{x},u:v}$ associated with the individual $y_{\tau:t}$, which can be retrieved after obtaining $\delta(T)$.

The full search for the optimal training segments can be accelerated by dynamically pruning those training segments producing small likelihoods $p(y_{\tau:t}|m_{\mathbf{x},u:v})$ for a given testing segment.

5. EXPERIMENTAL STUDIES

The TIMIT database was used in the study. We have evaluated the new approach in two aspects: (1) the ability to identify matching segments between the training and testing sentences, and (2) the applicability to speech recognition. For the first aspect, we considered the synthesis of the testing sentences using the matching training segments. For the second aspect, we considered the use of the phonetic transcripts associated with the matching training segments to label the testing segments/sentences.

The speech was divided into frames of 20 ms with a frame period of 10 ms. We used a 39-dimensional feature vector for each frame, consisting of cepstral coefficients $C_1 - C_{12}$ and normalized log energy, appended with their first- and second-order delta coefficients. The TIMIT database contains 3696 training sentences from 462 speakers (326 male, 136 female). We pooled all the training sentences together and trained a universal GMM, (1), with 4096 Gaussian components each with a diagonal covariance matrix. Then we built a model, (2), for each training sentence. Since all the 3696 training sentence models were built on the same 4096 Gaussian components, our system has a low parametric complexity in comparison to most recognition systems that use context-dependent acoustic-phonetic models. We performed the experiments on the *core* test set, consisting of 192 sentences from 24 untrained speakers (16 male, 8 female). While the new approach is capable of detecting matching segments with arbitrary lengths up to the complete sentences, in the experiments we restricted the search for the matching segments to a maximum length of 50 frames, or 510 ms, to reduce the amount of computation.

The new system segmented the 192 testing sentences, with a total of 58,015 frames, into 2,541 segments each matched by a segment found from the training sentences with optimal continuity. Fig. 1 shows the distribution of the length of these 2,541 optimal segments, with an average segment length of 22.8 frames. Fig. 2 shows the segment length in the number of phones spanned by the optimal segments, with an average length of 3.8 phones. Table 1 presents two specific examples for segmenting and recognizing two sentences, one from a male speaker (shorter) and the other from a female speaker (longer), labelled using the 61 phonemic/phonetic symbols used in TIMIT. Shown in the table are the optimal segmentation and the recognized phonetic transcript for each segment, compared to the reference transcript. In the table, we optionally place a sign '–' to the left/right of a segment, to indicate that the segment only contains part of the phones located at the borders of the segment. As each pair of the variable-length training and testing segments were compared as two whole units, the new system extracted the temporal dynamics over phonetic sequences. However, we notice that the system has rarely found the matching segments at the exact phonetic boundaries, which may indicate that phonetic boundaries are not stable islands for separating speech sounds.

To gain an idea about the phone recognition accuracy by the new system, we summarized the phone accuracy between the matching segments, as illustrated in Table 1. In the summarization, the incomplete phones at the borders between adjacent segments were treated as optional and mutually replaceable, as a way of modeling the uncertainty of the partial phones at the borders of segments. Following convention, we differentiated 39 phone classes. Table 2 presents the results, with comparisons to some of the previous best results reported on the core test set. The new system has not yet been comparable to other state-of-the-art systems but note that it has a much

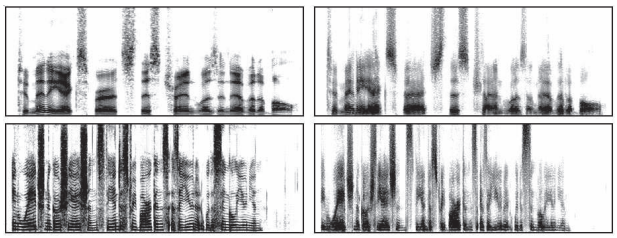

Fig. 3. Synthesis of two testing sentences by concatenating matching training segments with optimal continuities (left: original, right: synthesized).

lower parametric complexity, using only 4096 diagonal Gaussians.

We can synthesize the testing sentences by concatenating the Gaussians corresponding to the optimal matching training segments. Fig. 3 shows examples for synthesizing the two sentences in Table 1. As the new approach seeks maximum acoustic continuity, the vast majority of the synthesized sentences well resemble the original ones both naturally and with good word resolution, despite the difference in phonetic transcripts between some of the matching segments.

6. CONCLUSIONS

This paper described a new approach for modeling, segmenting and recognizing speech. The new approach consists of a sentence model to represent up to sentence-long temporal dynamics in the training data, and an algorithm to identify the matching segments with large continuities between the training and testing sentences. Recognition is performed by combining the longest matching segments from the training sentences. This should improve the separation of speech by exploiting their differences in long-range temporal dynamics. Preliminary experiments on the TIMIT database have shown encouraging results for identifying matching speech segments given the very low parametric complexity of the new approach.

7. REFERENCES

[1] M. Ostendorf, *et al*., "From HMM's to segment models: a unified view of stochastic modeling for speech recognition," IEEE Trans. Speech Audio Proc., vol. 4, pp. 360-378, 1996.

[2] J. R. Glass, "A probabilistic framework for segment-based speech recognition," Computer Speech and Language, vol. 17, pp. 137-152, 2003.

[3] V. Digalakis, J. Rohlicek, and M. Ostendorf, "A dynamical system approach to continuous speech recognition," IEEE Trans. Speech Audio Proc., vol. 1, pp. 431-442, 1993.

[4] R. Togneri and L. Deng, "Joint state and parameter estimation for a target-directed nonlinear dynamic system model," IEEE Trans. Signal Proc., vol. 51, pp. 3061-3070, 2003.

[5] F. Frankel and S. King, "Speech recognition using linear dynamic models," IEEE Trans. Speech Audio Proc., vol. 15, pp. 246-256, 2007.

[6] M. De Wachter, K. Demuynck, D. Van Compernolle, and P. Wambacq, "Data driven example based continuous speech recognition," Eurospeech'2003, pp. 1133-1136.

[7] S. Axelrod and B. Maison, "Combination of hidden Markov models with dynamic time warping for speech recognition," ICASSP'2004, pp. 173-176.

Table 1. Segmentation and recognition of two testing sentences: "To many experts, this trend was inevitable" (upper panel), and "In wage negotiations, the industry bargains as a unit with a single union" (lower panel). A '–' is used to indicate the phones that are only partly included in the segments.

Testing segment frame range	Matching training segment transcript	Reference transcript
0 13	–pau–	h#–
13 36	t ix m eh–	–h# t uh m eh–
36 72	–ih n iy eh kcl k–	–eh nx iy eh kcl k–
72 108	–s pcl p eh–	–k s pcl p er–
108 132	–ix tcl ch dcl–	–er tcl t s dh–
132 165	–pcl p ix s tcl t–	–dh ih s tcl t–
165 187	–z aa n–	–t r eh n dcl–
187 217	w ax z ah n–	dcl w ah z ih n–
217 245	nx eh v ix–	n eh v ax–
245 295	–ix dx ax bcl b el h#–	–ax dx ax bcl b el h#–
0 21	–h# ih–	h# q ih n–
21 33	–en w–	–n w–
33 56	–w ey–	–w ey–
56 74	–ih dcl jh en–	–ey dcl jh n–
74 83	–n ix–	n ix gcl–
83 91	–iy gcl g–	gcl g ow–
91 111	–ax sh–	ow sh–
111 155	–s iy ey sh epi en–	–sh iy ey sh ix–
155 190	–ao n s tcl t iy–	–ix n s epi dh iy–
190 214	iy eh n dcl d ae–	–iy ih n dcl d ix–
214 238	–r ax s t r ey–	–ix s tcl t r iy–
238 253	–n bcl b ay–	–iy bcl b aa–
253 269	–aa r–	–aa r gcl–
269 289	–eh kcl k en q pau–	gcl g ih n–
289 321	–eh tcl s q ih z ax–	–n z q eh z ey–
321 334	–iy y–	ey y–
334 361	–iy n ih z–	–y ux nx ih q–
361 372	–ey epi w–	–q w–
372 386	–ih dx ix z–	–w ix dh ix s–
386 422	–z ix n dh ax l–	–s ih ng gcl g el–
422 443	–l iy ix n–	–el y ux–
443 460	–iy ng ih–	–ux n y ih–
460 507	–eh m pau–	–ih n h#–
507 531	–h#–	–h#–
531 551	–h#–	–h#–

Table 2. Phone recognition on the TIMIT core test set.

Method	Phone error rate (%)
The new system	28.4
Triphone CD HMM [9]	27.1
Bayesian triphone HMM [10]	25.6
Segmental recognizer [2]	24.4

[8] G. Aradilla, J. Vepa, and H. Bourlard, "Improving speech recognition using a data-driven approach," Eurospeech'2005, pp. 3333-3336.

[9] L. Lamel and J. L. Gauvian, "High perfromance speaker-independent phone recognition using CDHMM," Eurospeech'1993, pp. 121-124.

[10] Ji Ming and F. J. Smith, "A Bayesian triphone model," Computer Speech and Language, vol. 13 pp. 195-206, 1999.