Segment-Based Recognition on the PhoneBook Task: Initial Results and Observations on Duration Modeling

Karen Livescu and James Glass

Spoken Language Systems Group
MIT Laboratory for Computer Science
Cambridge, Massachusetts 02139 USA
{klivescu, glass}@mit.edu

Abstract

This paper describes preliminary recognition experiments on PhoneBook [1], a corpus of isolated, telephone-bandwidth, read words from a large (almost 8,000-word) vocabulary. We have chosen this corpus as a testbed for experiments on the language model-independent parts of a segment-based recognizer. We present results showing that a segment-based recognizer performs well on this task, and that a simple Gaussian mixture phone duration model significantly reduces the error rate. We compare context-independent, stress-dependent, and word position-dependent duration models and obtain relative error rate reductions of up to 12% on the test set. Finally, we make some observations regarding the effects of stress and word position in this isolated-word task and discuss our plans for further research using PhoneBook.

1. Introduction

The work described in this paper was motivated by a desire to study aspects of automatic speech recognition related to the acoustic-lexical interface and segmentation in a segment-based recognizer. The PhoneBook database [1], a large-vocabulary isolated-word corpus, seems particularly well-suited to this type of investigation: the isolated-word task allows us to ignore the effects of a language model and to explore computationally intensive low-level modeling, while the large vocabulary makes the task sufficiently challenging. This paper describes our baseline system and initial experiments with duration models.

The following sections present the segment-based recognition framework and the interpretation of a duration model as a segmentation probability; describe the PhoneBook corpus, our baseline recognizer, and the initial duration models we have investigated; and interpret our results and suggest future directions.

2. Segment-based recognition framework

We begin by reviewing the probabilistic formulation of the speech recognition problem, which is used in our segment-based recognizer [2]. We present the general case in which an utterance may contain multiple words, although the experiments we discuss are limited to an isolated-word task. In the general formulation, the goal is to determine the most likely string of words $W^* = w_1, \ldots, w_N$ given the set of acoustic observations A, that is $W^* = \arg\max_W P(W|A)$, where W ranges over all possible word strings. Since a single word string can have multiple realizations as strings of sub-word units (e.g. phones) $U = u_1, \ldots, u_L$ with different segmentations S, this becomes

$$W^* = \arg\max_W \sum_{\forall U,S} P(W, U, S|A),$$

where U ranges over all possible pronunciations of the word string W and S ranges over all possible segmentations (i.e. locations of phone boundaries, or any other definition of a segmentation). We then make the conventional assumption that, given a word string W, there exist both an optimal pronunciation U^* and an optimal segmentation S^*, which are much more likely than any other (U, S) combination, so that we may replace the summation by a maximization:

$$\{W^*, U^*, S^*\} = \arg\max_{W,U,S} P(W, U, S|A).$$

Applying Bayes' rule several times, we can rewrite this as:

$$\{W^*, U^*, S^*\} = \arg\max_{W,U,S} P(A|W, U, S)P(S|U, W)P(U|W)P(W),$$

where the first term corresponds to the acoustic model, the second to the segmentation probability, the third to the pronunciation model, and the last to the language model.

In a typical frame-based recognizer, an HMM is used to jointly model the acoustic and segmentation terms (the latter being represented by the state transition probabilities). In a segment-based approach, on the other hand, the acoustic and segmentation terms are modeled separately. When no explicit model is used for the $P(S|U, W)$ term, it is in effect assumed to be a constant for any allowable segmentation of U and zero for any other segmentation.

The segmentation term, however, can be interpreted as a duration model as follows. Define a segmentation to be a list of the phone boundaries in that segmentation. That is, $S = t_1, \ldots, t_L$ denotes that the first phone extends from time 0 to t_1, the second extends from t_1 to t_2, and so on until the L^{th} phone. This is equivalent to the duration of the first phone being t_1, that of the second being $t_2 - t_1$, and so on. Using the notation s^d to denote the event "segment s has duration d", we have that

$$\begin{aligned} P(S = t_1, \ldots, t_L|U, W) &= P(s_1^{t_1}, s_2^{t_2 - t_1}, \ldots, s_L^{t_L - t_{L-1}}|U, W) \\ &= \prod_{l=1}^{L} P(s_l^{t_l - t_{l-1}}|U, W, s_1^{t_1}, s_2^{t_2 - t_1}, \ldots, s_{l-1}^{t_{l-1} - t_{l-2}}). \end{aligned}$$

This material is based upon work supported by the NSF under Grant No. IRI-9618731, and by DARPA under contract N66001-99-1-8904 monitored through the Naval Command, Control and Ocean Surveillance Center.

To be presented at the 7th European Conference on Speech Communication and Technology (EUROSPEECH), Aalborg, Denmark, September, 2001.

In the above we have considered only the case in which the number of segments in the segmentation is equal to the number of phones; for other cases the probability is zero (assuming that phones are not permitted to have zero duration). If we make the assumption that, given the unit string, each phone's duration is independent of the durations of the other phones, the conditioning statement in the last expression becomes (U, W) alone. This assumption roughly corresponds to ignoring the effects of speaking rate.

From this general duration model, we can obtain some simpler models by making various assumptions. For example, if we assume that each phone's duration depends only on the phone's identity and not on the identities of other phones or on the word string, we have the context-independent model

$$P(S|U,W) = \prod_{l=1}^{L} P(s_l^{t_l - t_{l-1}}|u_l).$$

Alternately, assuming that the dependence on U and W is captured solely by the current phone's stress and/or position within the word or utterance, in addition to the phone's identity, we have

$$P(S|U,W) = \prod_{l=1}^{L} P(s_l^{t_l - t_{l-1}}|u_l, str(u_l), pos(u_l)),$$

where $str(u_l)$ is the stress level of u_l and $pos(u_l)$ is some function of the position of u_l within an utterance. Effects such as lengthening of stressed vowels and of phrase-final segments are well-known [3] and have been used successfully to predict phone durations [4]. We discuss the use of such context-dependent duration models, as well as of a context-independent one, in Section 4.2.

The above interpretation of the duration model is similar to that in [5]. Duration models with various types of distribution have been attempted (e.g. Poisson [6], Gamma [7], and Gaussian mixtures [8]) and used in both frame-based [5] and segment-based [9, 10] recognizers. However, duration models are particularly natural to incorporate in a segment-based recognizer, since the search inherently considers entire segments and therefore requires no modification other than an additional score on each segment.

3. The PhoneBook database

The PhoneBook database contains approximately 92,000 utterances of isolated words read over the phone by native speakers of various dialects of American English. The vocabulary consists of almost 8,000 words of varying lengths (e.g. *aced, acoustically, winfrey*) designed to cover as many phonemic contexts as possible. The inclusion of many confusable word pairs (e.g. *scheduled/schedules, bulls/bolts*) makes the task challenging, especially when recognizing with the entire vocabulary. PhoneBook has typically been used for investigations either into tasks where the training and test vocabularies are different [11] or into new types of acoustic modeling and representation (e.g. modeling additional dependencies in [12] and [13], and articulatory state models in [14]).

We use the same breakdown of the database into training, development, and test sets as defined in [11]. There is no overlap between speakers or words in the different sets. Two training sets are defined; the "small" training set contains about 20,000 utterances and the "large" set contains about 80,000 utterances. The development and test sets contain about 7,000 utterances each. Most published results involve training on the 20k training set and classifying each word in the test or development set from among a list of words ranging in size between about 75 and about 600. We have chosen to train on the 80k training set in order to gain more freedom to train complex models, and to test with the full (8,000-word) vocabulary in order to increase the difficulty of the task. For comparison with the literature, however, we also report our baseline results when training on the 20k training set and testing with a 600-word vocabulary.

4. Experiments

4.1. Baseline recognizer

The baseline recognizer uses the SUMMIT segment-based system [2], [15]. It begins with a segmentation step that hypothesizes landmarks, or locations in the waveform at which phonetic boundaries are likely to occur, using an acoustic change criterion [10]. This step creates a graph of allowable segmentations and therefore limits the number of segmentations that must be considered in the search. Landmark features, consisting of MFCC averages over several regions around each landmark, are then extracted and scored with phonetic diphone acoustic models, as in [15]. A diphone can correspond to a landmark at the boundary between two phones or to a phone-internal landmark detected by the segmentation algorithm. Segment models, which model the regions between landmarks, can also be used in SUMMIT, but we do not currently use them for PhoneBook experiments.

Baseforms for each word are taken from the Pronlex dictionary [16] whenever possible, and from the dictionary provided with PhoneBook for words not appearing in Pronlex. Pronunciation rules such as flapping, palatalization, etc. are then applied and stress is removed to obtain a pronunciation graph for each word using a detailed set of 69 segment labels (66 phones plus silence models). We currently assume that all resulting pronunciations of a word are equally likely. Since the task is isolated-word and each word is equally likely, the language model is also trivial.

The SUMMIT recognizer uses a Viterbi training paradigm, which was seeded with phonetic alignments obtained using existing telephone-speech acoustic models from conversational domains [15]. Since the amount of training data is insufficient to train a model for each diphone occurring in the vocabulary, the diphones were then clustered using top-down decision tree-based clustering on the PhoneBook training data. We note that using top-down clustering is important in PhoneBook, as in any domain in which the training and test vocabularies are different; bottom-up clustering would be unable to assign a class to a diphone appearing in a test set but not in the training set.

The first line of Table 1 shows the error rates obtained with this baseline system on the 600- and 8,000-word tasks when training on the 20k training set. We note that this recognizer achieves a lower error rate than any that we have seen in the literature on the 600-word task. Table 2 shows published results on the 600-word task using the 20k training set.

Having established that the recognizer performs competitively, we switched to training on the 80k training set and testing on the 8,000-word task. At this point, we also switched to a phone set and pronunciation graphs with binary vowel stress information (where "stressed" vowels are those with primary stress and "unstressed" vowels either are reduced or have secondary stress), as well as tags indicating word-final vowels, for subsequent experiments with context-dependent duration mod-

Training set	# params	600-wd ER	8,000-wd ER
20k	627k	3.6	13.6
80k	1.55M	2.3	9.9

Table 1: *Error rates (ER, in %) of the baseline recognizer on the PhoneBook test set.*

Reference	Description, # params	ER
Dupont et al. [11]	hybrid HMM/ANN, 166k	5.3
Bilmes [13]	HMM or Dynamic Bayesian Multinet, ∼200k	5.6
Richardson et al. [14]	HMM + Hidden Articulator MM, 458k	4.17

Table 2: *Published test set error rates (ER, in %) on the 600-word PhoneBook task using the 20k training set.*

els. The pronunciation rules and diphone models, however, remain the same; the stressed and unstressed, and word-final and non-final, versions of the same vowel are classed in the same diphone class and are subject to the same rules as before. The performance of this recognizer is shown in the second line of Table 1.

4.2. Duration models

While examining the paths chosen by the baseline recognizer, we noted that many misrecognized utterances had unusually long or short phones and conjectured that a duration model would be helpful. Figures 1 and 2 show two example decoded utterances demonstrating the types of duration errors that we have noticed in the baseline recognizer.

We used a Gaussian mixture model for the phone duration density, since this model is simple and versatile; a discrete distribution is inappropriate in our case because the segmentation algorithm does not produce purely discrete duration values. In fact, we modeled the log duration rather than the duration itself, since the former is more conducive to modeling with a Gaussian mixture distribution. The number of Gaussians per model was determined for each class based on the number of training examples, up to a certain maximum number of Gaussians. From qualitative observation of the duration distributions in the training set, we set the maximum number of Gaussians to 5, in order to allow a good fit to the data while avoiding overfitting. We also found it beneficial to scale the duration scores relative to the diphone scores to account for the difference in their scales, as well as to add an offset to each segment's score to account for the recognizer's preference for longer or shorter phone sequences; the values of these parameters were set based on development set performance.

We began with a context-independent duration model, that is, $P(s_l^{t_l - t_{l-1}} | U, W) = P(s_l^{t_l - t_{l-1}} | u_l)$. In addition, we also trained duration models with three kinds of context dependence. In the first condition (the "stress-dependent" condition), separate models are trained for stressed and unstressed vowels. However, while stressed vowels usually tend to be longer than unstressed ones, we found that for some vowel classes in the PhoneBook training set, the unstressed version tends to be *longer* than the stressed one. This is less surprising once we realize that, in this training set, unstressed vowels are about 3.5 times more likely to be word-final than their stressed counterparts; and since the utterances are isolated words, word-final

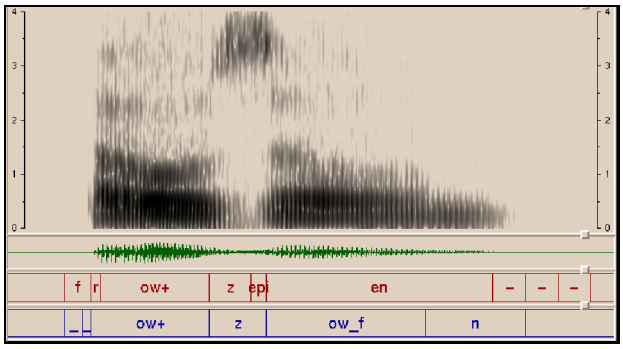

Figure 1: *Spectrogram of an utterance incorrectly recognized by the baseline recognizer but correctly recognized when a duration model is added. The top transcription shows the hypothesis produced by the baseline recognizer, corresponding to the word "frozen"; note the unusually short* [r] *(∼15 ms) and unusually long syllabic* [n], [en] *(∼380 ms). The bottom transcription shows the hypothesis produced with a duration model, corresponding to the correct word "ozone" (the label* [ow_f] *indicates a word-final* [ow]).

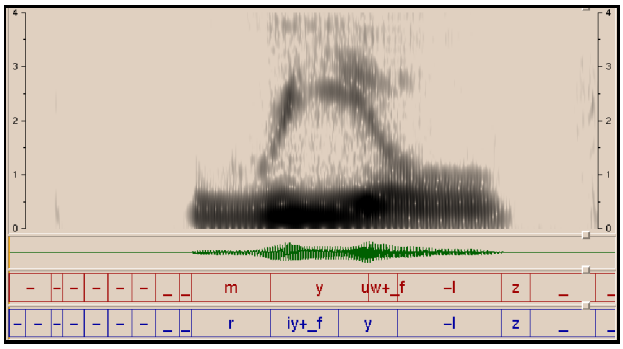

Figure 2: *Spectrogram of an utterance incorrectly recognized by the baseline recognizer but correctly recognized when a duration model is added. The top transcription shows the hypothesis produced by the baseline recognizer, corresponding to the word "mules"; note the long* [y] *(∼170 ms) and short* [uw+_f] *(stressed* [uw] *in word-final position, ∼50 ms). The bottom transcription shows the hypothesis produced with a duration model, corresponding to the correct word "reels".*

vowels are actually subject to *utterance*-final lengthening. We therefore trained a second set of context-dependent models that take into account whether or not a vowel is word-final. Finally, we trained a set of models with both stress and word-position dependence.

Table 3 summarizes the results of using all four duration models on the development and test sets. All of the duration models produce a statistically significant reduction in error rate over the baseline (using a McNemar test at a significance level of 0.01). On the development set, the four duration models are statistically equivalent. On the test set, the models with position dependence and those with both stress and position dependence perform significantly better than the context-independent and stress-only models. The test set results, therefore, support our hypothesis that word-final effects taken alone are more important than stress in modeling vowel duration in PhoneBook, although the development set results are less supportive of this claim. The best-performing models achieve a 12% relative error rate reduction on the test set. It is noteworthy that all of

Model	# params	dev ER	test ER
baseline	1.55M	10.5	9.9
context-ind't dur	+ 954	9.4	9.2
stress-dep't dur	+ 1179	9.4	9.1
position-dep't dur	+ 1179	9.3	8.7
stress- & position-dep't dur	+ 1614	9.4	8.7

Table 3: *Effect of duration models on the error rates for the development and test sets. The number of parameters for each duration model refers to the additional parameters needed by that model.*

these gains are obtained with a very small cost in the number of additional parameters to be estimated, as shown in the table.

Figures 1 and 2, showing the utterances that were previously poorly decoded with the baseline recognizer, also include the corresponding hypotheses when using a duration model (in this case, the context-independent model). In both cases, the recognizer using the duration model is able to correctly decode the utterances.

5. Future work

The results we have presented represent our preliminary investigation into the PhoneBook database as a testbed for experiments with segment-based recognition. These initial results demonstrate both the effectiveness of the segment-based approach on the PhoneBook task and the usefulness and ease of incorporating an explicit duration model into the recognition search. We plan to continue to use PhoneBook to explore duration modeling and other aspects of recognition on the acoustic/lexical level. Examples of additional issues we would like to explore using PhoneBook include articulatory feature and phonetic class representations for lexical access.

In the area of duration modeling, we have not taken into account many factors that can contribute to the performance of the models, such as speaking rate variations, phonetic context, and position effects other than utterance-final lengthening. While the models we have used thus far produce significant performance gains at a very small cost, we are interested in investigating the benefits of a more realistic model of duration. In order to better model contextual effects, we plan to incorporate a hierarchical duration model previously developed in our group [9] as a post-processing step. We have found that only a very shallow N-best list would be required to ensure that the correct hypothesis is present. Specifically, for the baseline recognizer, almost half of the errors on the development set correspond to the correct hypothesis being ranked second (4.7% out of a total error rate of 10.5%), and an N-best list with 10 hypotheses contains the correct hypothesis 98.7% of the time. When using a context-independent duration model, errors in which the correct hypothesis is ranked second account for 4.3% of the 9.4% error rate, and the correct hypothesis is in the 10-best list 99.0% of the time. This makes a rescoring approach attractive for this task.

Finally, we have observed that the distributions of phone durations can be erratic, due to both the inherently quantized nature of frame durations and certain intrinsic characteristics of the segmentation algorithm. This suggests that a different kind of distribution, possibly a semi-discrete one, may be preferable.

6. Acknowledgments

We would like to thank Jon Yi for his help in syllabifying the PhoneBook pronunciations.

7. References

[1] J. Pitrelli, C. Fong, S. Wong, J. Spitz, and H. Leung, "PhoneBook: A Phonetically-Rich Isolated-Word Telephone-Speech Database," in *Proc. ICASSP*, Detroit, Michigan, 1995.

[2] J. Glass, J. Chang, and M. McCandless, "A Probabilistic Framework for Feature-Based Speech Recognition," in *Proc. ICSLP*, Philadelphia, Pennsylvania, 1996.

[3] D. H. Klatt, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *J. Acoust. Soc. Amer.* 59 (5): 1208–1221, 1976.

[4] J. F. Pitrelli, *Hierarchical Modeling of Phoneme Duration: Application to Speech Recognition*. Ph.D. Thesis, MIT Dept. of Elec. Eng. and Comp. Sci., 1990.

[5] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, "From HMM's to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition," *IEEE Trans. Speech and Audio Processing*, 4 (5): 360–378, 1996.

[6] M. Russell and R. Moore, "Explicit Modeling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition," in *Proc. ICASSP*, Tampa, Florida, 1985.

[7] S. E. Levinson, "Continuously Variable Duration Hidden Markov Models for Speech Analysis," in *Proc. ICASSP*, Tokyo, Japan, 1986.

[8] Y. Deng, T. Huang, and B. Xu, "Towards High Performance Continuous Mandarin Digit String Recognition," in *Proc. ICSLP*, Beijing, China, 2000.

[9] G. Chung and S. Seneff, "A hierarchical duration model for speech recognition based on the ANGIE framework," *Speech Communication*, 27:113–134, 1999.

[10] V. Zue, J. Glass, D. Goodine, M. Phillips, and S. Seneff, "The SUMMIT Speech Recognition System: Phonological Modelling and Lexical Access," in *Proc. ICASSP*, Albuquerque, New Mexico, 1990.

[11] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J.-M. Boite, "Hybrid HMM/ANN Systems for Training Independent Tasks: Experiments on PhoneBook and Related Improvements," in *Proc. ICASSP*, Munich, Germany, 1997.

[12] G. Zweig and S. J. Russell, "Speech Recognition with Dynamic Bayesian Networks," in *Proc. AAAI*, Madison, Wisconsin, 1998.

[13] J. A. Bilmes, "Dynamic Bayesian Multinets," in *Proc. 16^{th} Conf. on Uncertainty in Artificial Intelligence*, Stanford, California, 2000.

[14] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-Articulator Markov Models: Performance Improvements and Robustness to Noise," in *Proc. ICSLP*, Beijing, China, 2000.

[15] J. R. Glass, T. J. Hazen, and I. L. Hetherington, "Real-Time Telephone-Based Speech Recognition in the JUPITER Domain," in *Proc. ICASSP*, Phoenix, Arizona, 1999.

[16] Linguistic Data Consortium, http://www.ldc.upenn.edu/.