S11.6

Modeling Acoustic-phonetic Detail in an HMM-based Large Vocabulary Speech Recognizer[1]

L. Deng, M. Lennig[2], V.N. Gupta[2], and P. Mermelstein[2]
INRS–Télécommunications
3 Place du Commerce
Montreal, Québec, Canada H3E 1H6

Abstract

This paper is focused on the acoustic recognizer of the INRS–Télécommunications 60,000-word vocabulary isolated-word recognition system. The task of the acoustic recognizer is to generate a list of word hypotheses and their likelihoods based on the acoustic data for each input word. Incorporation of detailed acoustic-phonetic knowledge into the acoustic recognizer is important to achieve a high level of recognition accuracy. We report two sets of experiments where such knowledge is incorporated into the hidden Markov models (HMMs) employed during recognition. In the first set, we employ vowel durational properties in the HMMs. In the second set, we model word-initial and word-final stop consonants as a sequence of context-dependent sub-phonemes. The performance of the recognizer is significantly improved by appropriate utilization of the context-dependent vowel-duration information and the context-dependent microsegmental properties of stop consonants.

I. Introduction

The objective of the work reported here is to maximize the acoustic recognition accuracy of a large-vocabulary recognizer for spoken isolated words. This speaker-dependent recognizer of words from a 60,000-word English vocabulary is based on phoneme representation by hidden Markov models (HMMs) (Gupta et al., 1987, Jelinek, 1976, Levinson et al, 1983). The 60,000 words include many highly confusable words, forcing maximal exploitation of the detailed acoustic-phonetic information available. We have conducted two sets of experiments where such information is incorporated into the HMMs.

First, we explore the use of durational differences in a vowel in different phonetic contexts. Each of the vowel phonemes is divided into three allophones, each corresponding to a largely distinctive range of vowel durations. Each vowel's HMM incorporates three separate sets of transition probabilities, one for each of the allophones. During recognition, an appropriate set of transition probabilities, according to the phonetic context where the vowel occurs, is selected to score the observation sequence.

Secondly, we focus on improving the poor discrimination manifested by the speech recognition system on the six stop consonants in CVC-type words. The improvement is obtained by modeling stop consonants as a sequence of microsegments: silence, voice bar, burst, and aspiration. The microsegmental models of burst and aspiration are conditioned on the adjacent vowel category: front vs. nonfront. These two contexts account for a large part of coarticulatory effects found in the burst and aspiration segments of stops.

In Sections II and III, we report the methodology and recognition results of these two sets of experiments, respectively. We draw our conclusions from these experiments in Section IV.

II. Incorporation of vowel duration into the recognizer

1 Development of vowel durational models

To provide a background for incorporating vowel duration into the HMMs, we examine the dependence of vowel duration on contextual factors in our training data set comprising 1102 isolated words spoken by a native American English speaker. To minimize complexity and to make best use of the limited number of word tokens available, we capture only the most distinct durational differences. In our experiments, the three contexts which lead to three distinctive vowel durations are as follows: (1) monosyllabic words with a voiced coda or in open syllables, (2) monosyllabic words with a voiceless coda, and (3) polysyllabic words. As an example, the duration of the vowel /o/ in word *code* (monosyllabic word with a voiced coda) is generally longer than that in word *coat* (monosyllabic word with a voiceless coda), which in turn is longer than that in word *coating* (polysyllabic word).

For each of the above contexts, the mean duration and standard deviation in the training data set are shown in Table 1. For all the vowels, the three allophones possess systematically different durations.

[1] This work was supported by the National Sciences and Engineering Research Council of Canada

[2] Also with Bell-Northern Research, Montréal, Canada

CH2561-9/88/0000-0509 $1.00 © 1988 IEEE

Vowel	voiced coda or open syllable in monosyllabic word	voiceless coda in monosyllabic word	in polysyllabic word
	mean (St.dev.)	mean (St.dev.)	mean (St.dev.)
aj	328 (60)	216 (25)	202 (47)
e	306 (58)	213 (18)	176 (34)
aw	294 (61)	260 (1)	252 (35)
ɔj	390 (29)	255 (25)	276 (69)
i	256 (60)	164 (25)	137 (61)
ɪ	218 (59)	156 (29)	104 (41)
ɛ	214 (36)	185 (24)	133 (38)
æ	303 (51)	234 (22)	159 (36)
ɑ	311 (47)	234 (29)	178 (33)
ʌ	266 (53)	174 (25)	129 (38)
u	294 (81)	186 (28)	168 (46)
ʊ	193 (26)	157 (21)	121 (37)
o	308 (68)	203 (24)	193 (57)
ɔ	262 (54)	203 (9)	167 (44)

Table 1. Durational statistics of vowels in milliseconds for training data (1102 words)

The above division of vowel tokens suggests that three separate HMMs be trained for each vowel in order to achieve more accurate modeling of vowel duration. In order to avoid a three-fold increase in the number of parameters to be estimated, we only train the transition probabilities separately for each allophone. The output observation probabilities (multivariate Gaussian) are jointly trained from all three allophones of the vowel. Tying of the output probabilities this way is reasonable since the spectral shape of the vowel does not differ appreciably among the three allophones. The tying results in only a 5%, instead of a three-fold, increase in model parameters to be estimated, ensuring robust estimation of the parameters of all three allophone models.

2 Overview of the recognition system

An overview of our large vocabulary isolated word recognizer, without the use of vowel durational models presented here, can be found in Gupta et al., 1987. Briefly, the recognition process consists of word endpoint detection, a fast-search algorithm to generate a list of the N most likely word choices, and the computation of exact likelihoods for these N choices. The durational HMM's of vowels presented in this paper are only introduced at the exact likelihood scoring stage.

Since the fast-search algorithm has already provided phoneme sequences for each hypothesized word, the phonetic context of the vowel is known during exact likelihood computation. This allows one of the three vowel allophones to be deterministically selected to score each observation.

The performance of the recognizer is evaluated using the following three measures. The first measure is the percentage of words correctly identified as the top word choice by the recognizer. The second measure is the average rank of the correct word in the ordered list of word hypotheses. The third measure is the average difference between the log likelihood of the correct word and the largest log likelihood among the incorrect words (i.e., the most confusable word).

3 Experimental evaluations

A comparison of the recognition performance using vowel HMMs with and without incorporating durational information is shown in Table 2 for a 782-word natural text test set, and in Table 3 for a 312-word CVC(V) test set. Use of the vowel durational HMMs (column B), compared with use of standard HMMs (column A), consistently improves the recognizer's performance for all three evaluation criteria for both the CVC(V) and the natural text sets. For the natural text set, the error rate is reduced by 15.4%, and for the CVC(V) set, it is reduced by 13.3%.

Performance measure	Natural text (782 words)		
	A	B	C
Percent Correct	84.4	86.8	79.4
Average Rank	1.45	1.31	1.71
Ave.Diff.of Log Scores	36.4	40.9	35.9

Table 2. Comparison of recognition performance with and without vowel durational models for natural-text data set. Column A: one HMM per vowel; Column B: three HMMs per vowel with tied output probabilities; Column C: three HMMs per vowel with separate output probabilities.

Performance measure	CVC(V) set (312 words)		
	A	B	C
Percent Correct	63.8	68.6	62.0
Average Rank	2.58	2.30	2.80
Ave.Diff.of Log Scores	6.2	11.0	6.1

Table 3. Comparison of recognition performance with and without vowel durational models for CVC(V) test set.

We should emphasize that the vowel durational models improve recognition only when the model robustness is maintained by tying the output observation probabilities for all three allophone models. To illustrate this, an experiment was performed where the output observation probabilities were not tied. The recognition performance, shown in columns C of Tables 2 and 3, is significantly poorer.

Test Words		Top Choice Using Standard HMMs	Top Choice Using Duration HMMs
a	(√)	[pei]	[e]
are	(√)	[awr]	[ɑr]
bid	(√)	[bid]	[bɪd]
but	(√)	[bajt]	[bʌt]
deacon	(√)	[pɪkɪn]	[dikən]
dice	(√)	[dajz]	[dajs]
digger	(√)	[ditr]	[dɪgr]
fan	(×)	[fæn]	[fɛn]
future	(√)	[titr]	[fjutʃr]
gave	(×)	[gev]	[gɪv]
have	(√)	[hav]	[hæv]
is	(√)	[iz]	[ɪz]
lax	(√)	[lɑks]	[læks]
lost	(√)	[lʌst]	[lɔst]
nineteen	(√)	[bajtɪŋ]	[najntin]
one	(√)	[wan]	[wʌn]
pride	(√)	[kajnd]	[prajd]
slight	(√)	[slaj]	[slajt]
than	(×)	[ðæn]	[ðɛn]
tried	(√)	[tʃrajəd]	[tʃrajd]
was	(√)	[wajz]	[wʌz]
with	(√)	[wɪdθ]	[wɪθ]
why	(√)	[wajni]	[waj]

Table 4. Examples of test words on which recognition errors were corrected (marked by √) or newly introduced (marked by ×).

To gain insight into the advantages of using the new vowel durational models, we examined the word recognition errors corrected or introduced by using the vowel durational allophones (Table 4). As can be seen, the vowel durational modeling provides two different types of advantages. First, incorporation of vowel duration into its HMM makes the model a more precise discriminator of the vowel itself, accounting for many error corrections such as [lɑks] to [læks], [bajt] to [bʌt], and so on. Secondly, vowel duration provides a cue to the context, namely the syllabic coda (voiced or voiceless) or the syllable type (open or closed). Thus, improved modeling of the vowel not only reduces vowel confusions, but also improves discrimination of the postvocalic consonants. Evidence for this is found in the errors corrected, namely, from [dajz] to [dajs], and from [wɪdθ] to [wɪθ]. Several errors due to weakly released voiceless stops have also been corrected by the vowel durational models, due to the cue provided by the vowel duration as to whether the vowel is in an open syllable or in a syllable closed by a voiceless consonant.

A similar experiment was performed with speech of a second male speaker. Use of the vowel durational models improved the recognition rate from 77.0% to 79.3%, a 10% reduction in the error rate.

III. Context-dependent microsegmental modeling of stop consonants

1 Development of stop consonant models

Use of one HMM for each stop consonant results in poor discrimination among stop consonants, especially in the consonant-vowel-consonant (CVC) environment. The word recognition error rate for CVC words is nearly three times as high as that for the natural language test sets. This is because use of a single HMM to represent each stop consonant prevents precise modeling of many of its useful phonetic properties.

Our strategy to improve the modeling of stops is to use their microsegmental properties. The basis for this strategy is that different stops and different allophones of the same stop often share common microsegments. Since these common microsegments generally do not provide discriminative information among different stops, they are combined and trained together. The modeling can then be focused on those microsegments which are essential for discrimination. The small number of such microsegments allows more tokens to be used for their training despite the limited amount of training data. The robustness gained from using more training data allows improved modeling of these information-bearing microsegments in different contexts.

The following nine HMMs are used to represent the six stop consonants in the experiments:

1. **voice bar**, whose tokens are taken from the three voiced stops in word-initial (when present) and word-final positions; a two-state HMM is used.
2. **silence**, whose tokens are taken from the three voiceless stops in word-final position; a two-state HMM is used.
3. **voiced stop aspiration**, whose tokens are taken from the three voiced stops in word-final position; a three-state HMM is used.
4. **/b/ burst**, tokens are from word-initial and word-final positions; a two-state HMM is used.
5. **/d/ burst**, tokens are from word-initial and word-final positions; a two-state HMM is used.
6. **/g/ burst**, tokens are from word-initial and word-final positions; a two-state HMM is used.
7. **/p/ burst+aspiration**,tokens are from word-initial and word-final positions; a four-state HMM is used.
8. **/t/ burst+aspiration**, tokens are from word-initial and word-final positions; a four-state HMM is used.

9. **/k/ burst+aspiration**, tokens are from word-initial and word-final positions; a four-state HMM is used.

Although the above models represent a large proportion of stop consonants, many word-medial stops (for example, flaps and glottal stops), cannot be unambiguously segmented. These word-medial stops contributed to significantly fewer errors than the stops in CVC-type words, hence they are not the major concern of this study. To represent them by HMMs, we simply use the traditional one-HMM-per-stop approach. That is, each word-medial stop is represented by one HMM trained from all the word-medial stop tokens.

Certain complications often arise in the microsegmental modeling of stops. First, the voice bar of a word-initial voiced stop may or may not be present. Secondly, word-final voiceless stops are sometimes not released. Even when present, a weak initial voice bar or a final burst may be deleted by the automatic word endpoint detection algorithm, which is based on loudness. Very low recognition scores would result if the models for these microsegments were used to score the missing microsegments. To avoid this, we allow a null transition in parallel with the HMM representing these microsegments (word-initial voice bar or word-final voiceless stop).

Incorporating the microsegmental models in recognition significantly improves the stop consonants' discriminability. Further improvements are obtained by creating context-dependent HMMs for the burst microsegment of each voiced stop, and for the burst+aspiration microsegment of each voiceless stop. The spectral properties of these stop microsegments are strongly affected by the adjacent vowel. The strongest coarticulation effects for these microsegments are seen in front vs. nonfront vowel contexts, thus these two vowel contexts are invoked in the context-dependent modeling of the microsegments.

2 Experimental evaluations

We compare four different ways of modeling stop consonants: one HMM per stop, three-HMMs per stop (one for word-initial, one for word-final, and one for word-medial stops), nine context-independent microsegmental HMMs, and 15 context-dependent microsegmental HMMs. The test data consist of 312 CVC or CVCV words, from one native American English speaker. The test set is selected to be highly confusable, many words differing by only one stop consonant (such as bit, bid, bib, big, bitter and bigger).

Table 5 compares the recognition results for the four different ways of modeling stop consonants. As can be seen from the table, modeling stop consonants using 15 context-dependent microsegmental HMMs produces the best recognition results for all three different performance measures. The context-independent microsegmental stop models, while significantly better than the one-HMM-per-stop system, are somewhat inferior to the three-HMMs-per-stop system.

Type of model for stop consonants	**Performance**		
	Percent Correct	Average Rank	Ave.Diff.of Log scores
Context-dept.microseg	80.4	1.61	20.7
Context-indep.microseg	75.6	1.96	14.7
Three HMMs per stop	76.6	1.77	18.7
One HMM per stop	68.6	2.30	11.0

Table 5. Performance comparison for four types of HMMs for stop consonants using a 312-word CVC(V) test set.

IV. Conclusions

Vowel duration and stop consonant's microsegmental properties represent important context-dependent acoustic-phonetic information, and have been incorporated into an HMM-based large vocabulary speech recognizer. Recognition results show that appropriate use of this information can lead to a significant improvement in recognition accuracy, when steps are taken to retain robustness in parameter extraction by allowing only a small increase in total number of parameters to be estimated.

To incorporate vowel duration into the HMM, we condition the state-transition probabilities of vowel HMMs on the context-specific class of vowel allophones, resulting in a 15% reduction in error rate. This reduction is statistically significant (based on 95% confidence interval).

Modeling a stop consonant as a sequence of microsegments improves recognition performance for CVC(V) words. Context-independent modeling of the stop microsegments reduced the error rate by 22% compared to the one-HMM-per-stop recognition system. Conditioning the burst and aspiration microsegments of stops on the adjacent front vs. nonfront vowel contexts results in an additional 16% reduction in the error rate.

References

V. Gupta, M. Lennig and P. Mermelstein, "Integration of acoustic information in a large vocabulary word recognizer," *Proceedings of ICASSP-87*, vol. 2, pp. 697–700, 1987.

F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532–556, 1976.

S.E. Levinson, L.R. Rabiner and M.M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *B.S.T.J.*, vol. 62, no. 4, pp. 1035–1074, 1983.