An 86,000-Word Recognizer Based on Phonemic Models

M. Lennig, V. Gupta, P. Kenny, P. Mermelstein, D. O'Shaughnessy

INRS–Télécommunications
3 Place du Commerce
Montreal, Canada H3E 1H6
(514) 765-7772

Abstract

We have developed an algorithm for the automatic conversion of dictated English sentences to written text, with essentially no restriction on the nature of the material dictated. We require that speakers undergo a short training session so that the system can adapt to their individual speaking characteristics and that they leave brief pauses between words. We have tested our algorithm extensively on an 86,000 word vocabulary (the largest of any such system in the world) using nine speakers and obtained word recognition rates on the order of 93%.

Introduction

Most speech recognition systems, research and commercial, impose severe restrictions on the vocabulary that may be used. For a system that aims to do speech-to-text conversion, this is a serious limitation since the speaker may be unable to express himself in his own words without leaving the vocabulary. From the outset we have worked with a very large vocabulary, based on the 60,000 words in Merriam Webster's Seventh New Collegiate Dictionary. We have augmented this number by 26,000 so that at present the probability of encountering a word not in the vocabulary in a text chosen at random from a newspaper, magazine or novel is less than 2% [25]. (More than 80% of out-of-vocabulary words are proper names.)

Our vocabulary is thus larger than that of any other English language speech-to-text system. IBM has a real-time isolated word recognizer with a vocabulary of 20,000 words [1] giving over 95% word recognition on an office correspondence task. The perplexity [16] of this task is about 200; the corresponding figure in our case is 700. There is only one speech recognition project in the world having a larger vocabulary than ours; it is being developed by IBM France [20] and it requires that the user speak in isolated syllable mode, a constraint which may be reasonable in French but which would be very unnatural in English.

Briefly, our approach to the problem of speech recognition is to apply the principle of maximum a posteriori probability (MAP) using a stochastic model for the speech data associated with an arbitrary string of words. The model has three components: (i) a *language model* which assigns prior probabilities to word strings, (ii) a *phonological component* which assigns phonetic transcriptions to words in the dictionary and (iii) an *acoustic-phonetic model* which calculates the likelihood of speech data for an arbitrary phonetic transcription.

Language Modeling

We have trained a trigram language model, which assigns a prior probability distribution to words in the vocabulary based on the previous two words uttered, on 60 million words of text consisting of 1 million words from the Brown Corpus [11], 14 million from *Hansard* (the record of House of Commons debates), 21 million from the *Globe and Mail* and 24 million from the Montreal *Gazette*.[1] Reliable estimation of trigram statistics for our vocabulary would require a corpus which is several orders of magnitude larger and drawn from much more heterogeneous sources but such a corpus is not available today. Nonetheless we have found that the trigram model is capable of correcting over 60% of the errors made by the acoustic component of our recognizer; in the case of words for which trigram statistics can be compiled from the training corpus, 90% of the errors are corrected.

Perhaps the simplest way of increasing recognition performance would be to increase the amount of training data for the language model. Although we are fortunate to have had access to a very large amount of data, we are still a long way from having a representative sample of contemporary written English. IBM has trained their language model using 200 million words of text. It seems that at least one billion words drawn from diverse sources are needed.

We have found that it is possible to compensate to some extent for the lack of training data by training

[1] We take this opportunity to acknowledge our debt to the *Globe and Mail*, to the *Gazette*, to G. & C. Merriam Co., to InfoGlobe, and to Infomart. Also, this work was supported by the Natural Sciences and Engineering Research Council of Canada.

parts-of-speech trigrams rather than word trigrams [10]. One of our graduate students has produced a Master's thesis which uses Markov modeling and the very detailed parts-of-speech tags with which the Brown Corpus is annotated to annotate new text automatically. We have also developed a syntactic parser which is capable of identifying over 30% of the recognition errors which occur after the trigram model [22].

The Phonological Component

In most cases Merriam Webster's Seventh New Collegiate Dictionary indicates only one pronunciation for each word. The transcriptions do not provide for phenomena such as consonant cluster reduction or epenthetic stops. Guided by acoustic recognition[2] errors, we have devised a comprehensive collection of context-dependent production rules which we use to derive a set of possible pronunciations for each word. This work is described in [26].

Acoustic-Phonetic Modeling

With the exception of /l/ and /r/, we represent each phoneme by a single hidden Markov model. The outstanding advantage of Markov modeling over other methods of speech recognition is that it provides a simple means of matching an arbitrary phonetic transcription with an utterance. However it suffers from several well-known drawbacks: HMMs fail to represent the dynamics of speech adequately since they treat successive frames as being essentially independent of each other; they cannot be made sensitive to context-dependent phonetic variation without greatly increasing the number of parameters to be estimated; they do not model phoneme durations in a realistic way. We have made substantial contributions to the literature on each of these problems. Our approach has been to increase the speech knowledge incorporated in our models without increasing the training requirements unduly. This has generally paid off in significant improvements in recognition performance.

We were one of the first groups to advocate the use of dynamic parameters, calculated by taking differences between feature vectors separated by a fixed time interval, and we have patented this idea. In [12] we introduced the idea of *multiple codebooks*, which enables vector quantization HMMs using both static and dynamic parameters to be trained using reasonable amounts of data. This idea has been adopted by several other researchers, notably Lee and Hon [19] and BBN. (We no longer use it ourselves since we found early on that multivariate Gaussian HMMs outperform vector quantization HMMs on our task and that the problem of undertraining is much less severe in the Gaussian case [6]).

An unfortunate consequence of using both static and dynamic parameters in a HMM is that the resulting model is a probability distribution on 'data' which satisfy no constraints relating static and dynamic parameters. (The model does not know how the dynamic parameters are calculated from the static parameters.) In the multivariate Gaussian case, it follows that the model is inconsistent in the sense that the totality of the data it can be presented with in training or recognition is assigned zero probability. This inconsistency led us to construct a new type of linear predictive HMM [18] which contains the static parameters HMM, the dynamic parameters HMM and the Poritz hidden filter model [23] as special cases.

In recognition tasks with a medium sized vocabulary (on the order of 1,000 words), the method of triphone modeling [24] has been found to be successful in addressing the problem of context-dependent phonetic variation. In its present form, this method cannot be scaled up to a recognition task as large as ours. (The number of triphones in our dictionary is more than 17,000; when triphones spanning word boundaries are counted as well, the number is much larger [15].) However we found that by constructing a collection of twenty five generalized-triphone models for each phoneme we were able to get a substantial improvement in recognition performance over unimodal phonemic HMMs (benchmark results) [5]. The generalized-triphone units were defined by means of a five way classification of left and right contexts for each phoneme[3]. We use the preceding phoneme class for a vowel and the following phoneme class for a consonant to construct one-sided HMMs (also called L/R-allophonic HMMs). In constructing two-sided allophonic HMMs (LR-allophonic HMMs) for each phoneme, a combination of the above five contexts in both left and right gives rise to 25 two-sided allophonic contexts. The first conclusion we can draw from Table I is that allophonic HMM's (columns 5–8) consistently outperform unimodal phonemic HMM's (columns 3–4). The difference in recognition accuracy is particularly noticeable with a large amount of training data (e.g., over 2,500 words). In this case, averaged over speakers CA and AM, L/R-allophonic HMM's reduce recognition errors by 18% when we use the uniform language model

[2] That is, recognition performed without the benefit of the language model

[3] For vowels, neighboring phonemes were classified as: (1) word boundary, breath noise, or /h/, (2) labial consonants, (3) apical consonants, (4) velar consonants, (5) vowels. For consonants, neighboring phonemes were classified as (1) word boundary or breath noise, (2) palatal vowels (including /j/), (3) rounded vowels (including /w/), (4) plain vowels, (5) consonants.

Speaker	**Train**	**Benchmark**		**L/R-alloph.**		**LR-alloph.**		**Mixtures**	
(test size)	**size**	**unif.**	**3-gram**	**unif.**	**3-gram**	**unif.**	**3-gram**	**unif.**	**3-gram**
CA (female) (1090 words)	717	32.1	19.4	30.0	16.5	30.3	17.5	30.3	19.5
	1532	29.8	15.0	25.0	11.5	21.9	12.0	19.0	10.0
	2347	29.4	13.2	24.5	10.9	19.2	9.6	13.9	5.8
	3098	29.2	13.2	24.0	11.0	17.3	8.7	14.0	6.0
	3880	29.3	13.3	23.9	10.9	17.0	8.1	14.1	6.0
AM(male) (698 wds)	1100	45.5	22.0	46.0	22.0	46.6	23.0	44.6	21.0
	2039	31.8	19.0	27.0	16.0	28.0	17.0	26.1	13.0
	2742	31.3	18.1	25.8	11.9	23.9	13.6	23.3	10.3
MA(fem.) (586 wds)	1600	21.0	9.6	16.9	8.4	16.2	10.4	13.8	7.5

Table I. Comparison of recognition error rates (in %) for the context-dependent allophonic HMM's (L/R-allophone and LR-allophone models) and the context-independent phonemic HMM's (unimodal (benchmark) and mixture models). Results for the uniform (unif.) and trigram (3-gram) language models are given separately.

and by 26% when we use the trigram language model. LR-allophonic HMM's reduce the error rate further, by 35% and 33%, respectively, for the two language models.

One of our most interesting discoveries was that we could obtain still better performance by training Gaussian mixture HMMs for each phoneme with 25 mixture components per state, using the mean vectors of the generalized triphone models as an initialization. Since the forward-backward calculations are notoriously computationally expensive for mixture models having large numbers of components, we had to devise a new variant of the Baum-Welch algorithm in order to train our system. We call it the *semi-relaxed* training algorithm [8]. It uses knowledge of the approximate location of segment boundaries to reduce the computation needed for training by 70% without sacrificing optimality. (For continuous speech, the computational savings will be larger still.) As can be seen from Table I (compare columns 7–8 to columns 9–10), the mixture HMMs outperform the LR-allophonic HMMs in almost every instance both with the uniform and the trigram language models.

The acoustic realization of stop consonants is highly variable, making them the most difficult phonemes to recognize. In general, they may be decomposed into quasi-stationary subsegments (microsegments) which can be classified crudely as silence, voice-bar, stop-burst and aspiration; the microsegments that actually occur in the realization of a given stop depend largely on its phonological context. We performed an experiment where we trained HMMs for several different types of microsegment (15 in all) and formulated context-dependent rules governing their incidence. We obtained a dramatic improvement in the acoustic recognition rate for CVC words. When tested on two speakers (see Table II), the error rate improved from 32.4% to 22.1% in one case and from 31.4% to 19.6% in the other [4].

Much of the information for recognizing stops (and other consonants) is contained in the formant transitions of adjacent vowels. It is not possible for us to take advantage of this fact directly since there are far more CV and VC pairs in our dictionary than can be covered in a training set of reasonable size. However, we have constructed a model for these transitional regions which we call a *state interpolation* HMM [17] and which can be trained using data that contains instances of every vowel and every consonant but not necessarily of every CV and VC pair. The state interpolation HMM models the signal in the transitional region by assuming that it can be fitted to a line segment in the feature parameter space joining a vowel steady-state vector to a consonant locus vector (the terminology is motivated by [3]); the remainder of the signal is modeled by consonant and vowel HMMs in the usual way. One steady-state vector is trained for each vowel and one locus vector for each consonant, so the model is quite robust. When tested

speaker	Percent Error (no language model)		
	one HMM	stop microsegments	
	per stop	context-indep.	context-dependent
spkrl	32.4%	26.0%	22.1%
spkr2	31.4%	24.4%	19.6%

Table II. Comparison of recognition error rates for 312 CVC(V) words using one model per stop, context-independent microsegment models, and context-dependent microsegment models. No language model is used.

on five speakers we found that this model gave improvements in acoustic recognition performance in every case; it also gives consistent improvements across a variety of feature parameter sets.

We have observed marked differences in the distribution of vowel durations in certain environments and we have found that this can be used to improve recognition performance by conditioning the transition probabilities (but not the output distributions) of the vowel HMMs on these environments [7]. We have performed recognition experiments where we distinguish three environments for each vowel: monosyllabic words with a voiceless coda, monosyllabic words having a voiceless or absent coda and polysyllabic words. This gave a 2% increase in acoustic recognition accuracy for both the speakers tested.

Many acoustic misrecognitions in our recognizer are due to phonemic hidden Markov models mapping to short segments of speech. When we force these models to map to larger segments corresponding to the observed minimum durations for the phonemes [14], then the likelihood of the incorrect phoneme sequences drops dramatically. This drop in the likelihood of the incorrect words results in significant reduction in the acoustic recognition error rate. Even in cases where acoustic recognition performance is unchanged, the likelihood of the correct word choice improves relative to the incorrect word choices, resulting in significant reduction in recognition error rate with the language model. On nine speakers, the error rate for acoustic recognition reduces from 18.6% to 17.3%, while the error rate with the language model reduces from 9.2% to 7.2%.

Overview of the Recognizer

Speech is sampled at 16 kHz and a 15-dimensional feature vector is computed every 10 ms using a 25 ms window. The feature vector consists of 7 mel-based cepstral coefficents [2] and 8 dynamic parameters calculated by taking cepstral differences over a 40 ms interval. (The zeroth order cepstral coefficent which contains the loudness information is not included in the static parameters but it is used in calculating the dynamic parameters.)

The first step in recognizing a word is to find its endpoints, which we do using a weighted spectral energy measure. In order to avoid searching the entire dictionary, we then attempt to 'recognize' the number of syllables in the word using a vector quantization HMM trained for this purpose, generating up to three hypotheses for the syllable count. The correct count is found in the hypothesis list 99.5% of the time.

For each of the hypothetical syllable counts we generate a list of up to 100 candidate phonetic transcriptions using crude forward-backward calculations and our *graph search* algorithm [13] to search a syllable network for transcriptions which are permitted by the lexicon. The exact likelihood of the speech data is then calculated for each of the candidate transcriptions using the acoustic-phonetic model. We thus obtain the acoustic match of the data with up to 300 words in the vocabulary (the number of words depends on the number of hypotheses for the syllable count). This list of candidate words is found to contain the correct word 96.5% of the time when phonemic duration constraints are not imposed on the search. In this case the search takes about two minutes to perform on a Mars-432 array processor. The percentage increases to 98% when the search is constrained to respect minimum durations. We have also found that the number of search errors can be reduced by using the language model to generate additional word hypotheses, but this increases recognition time by a factor of two so we do not use it.

At this point we have a lattice of acoustic matches for each of the words uttered by the speaker. The final step is to find the MAP word string by using the acoustic matches to perform an A* search [21] through the language model.

Word recognition rates on data collected from nine speakers are presented in Table III. For each speaker,

speaker	total words		acoustic recognition				errors after lang model	
			search errors		recog errors			
(sex)	training	test	no dur	dur	no dur	dur	no dur	dur
DS (m)	1900	451	4.9%	3.1%	24.0%	21.9%	14.4%	10.4%
AM (m)	2742	565	3.6%	1.8%	30.6%	31.0%	14.2%	12.2%
ML (m)	2000	596	1.7%	1.2%	14.5%	12.6%	6.7%	5.4%
JM (m)	1664	587	3.0%	3.0%	23.9%	22.7%	8.2%	7.8%
FS (m)	2322	1014	1.7%	1.3%	8.4%	7.5%	5.0%	3.7%
NM (f)	1299	967	4.0%	1.7%	19.4%	15.0%	11.0%	6.0%
CM (f)	2343	1090	3.5%	2.1%	16.9%	16.5%	8.9%	7.8%
MM (f)	2338	586	2.2%	1.4%	14.3%	14.3%	5.0%	3.6%
LM (f)	2353	863	3.8%	1.7%	23.7%	22.6%	12.1%	9.8%
Ave/Tot	2107	6719	3.1%	1.8%	18.6%	17.3%	9.2%	7.2%

Table III. Recognition error rates for nine speakers with and without duration constraints.

the number of word tokens used in training and testing are listed in the first two columns. The test data comprise 6,719 word tokens in all; recall that there are 86,000 words in the vocabulary.

Conclusions

Our objective was to develop an algorithm for speech-to-text conversion of English sentences spoken as isolated words from a very large vocabulary. We started with a vocabulary of 60,000 words but we found it necessary to increase this number to 86,000. Our initial recognizer used VQ-based HMMs, but we have since then switched to Gaussian mixture HMMs resulting in dramatic reduction in acoustic recognition errors. Imposing duration constraints on these HMMs has resulted in further reductions in acoustic recognition errors. We have shown that the trigram language model can be used effectively in our 86,000-word vocabulary recognizer, reducing the recognition errors by another 60%.

The recognition results show that we have acquired the capability to recognize words drawn from this much larger vocabulary with a degree of accuracy which is sufficient to warrant the commercial development of this technology once real-time implementation problems have been solved. Professor Jack Dennis of MIT has proposed a parallel architecture for HMM-based continuous speech recognition [9]. He estimates that decoding time can be decreased by a factor of at least 100 using a ‘parallel priority queue’. We have recently begun to explore this avenue.

References

1. Averbuch, A. et al., “Experiments with the Tangora 20,000 word speech recognizer,” Proc. 1987 IEEE International Conference on Acoustics, Speech, and Signal Processing, 701–704.

2. Davis, S.B., and Mermelstein, P., “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech, and Signal Processing,* **ASSP-28** (4), 357–365, 1980.

3. Delattre, P., Liberman, A.M., and Cooper, F.S., “Acoustic loci and transitional cues for consonants,” J. Acoust. Soc. Am. **27**, 769–774, 1955.

4. Deng, L., Lennig, M., and Mermelstein, P., “Modeling microsegments of stop consonants in a hidden Markov model based word recognizer.” J. Acoust. Soc. Am. **87** (6), 2738–2747, 1990.

5. Deng, L., Lennig, M., Seitz, F. and Mermelstein, P., “Large vocabulary word recognition using context-dependent allophonic hidden Markov models,” Computer Speech and Language, in press, 1990.

6. Deng, L., Kenny, P., Lennig, M., and Mermelstein, P., “Modeling acoustic transitions in

speech by state-interpolation hidden Markov models," IEEE Trans. on Acoustics, Speech, and Signal Processing, in press, 1990.

7. Deng, L., Lennig, M., and Mermelstein, P., "Use of vowel duration information in a large vocabulary word recognizer," J. Acoust. Soc. Am. **86** (2), 540–548, 1989.

8. Deng, L., Kenny, P., Lennig, M., Gupta, V., Seitz, F., and Mermelstein, P., "Phonemic hidden Markov models with continuous mixture output densities for large vocabulary word recognition," correspondence item, IEEE Trans. on Acoustics, Speech, and Signal Processing, in press, 1990.

9. Dennis, J., "Dataflow computation for artificial intelligence," *Parallel processing for supercomputers and artificial intelligence*, Edited by Kai Hwang and Doug DeGroot, McGraw Hill.

10. Dumouchel, P., Gupta, V., Lennig, M., and Mermelstein, P., "Three probabilistic language models for a large-vocabulary speech recognizer," Proc. 1988 IEEE International Conference on Acoustics, Speech and Signal Processing, 513–516.

11. Francis, W.N., and Kucera, H., "Manual of information to accompany a standard sample of present-day edited American English, for use with digital computers," Department of Linguistics, Brown University, 1979.

12. Gupta, V., Lennig, M., and Mermelstein, P., "Integration of acoustic information in a large vocabulary word recognizer," Proc. 1987 IEEE International Conference on Acoustics, Speech and Signal Processing, 697–700.

13. Gupta, V., Lennig, M., and Mermelstein, P., "Fast search strategy in a large vocabulary word recognizer," J. Acoust. Soc. Am. **84**(6), 2007–2017, 1988.

14. Gupta, V., Lennig, M., Mermelstein, P., Kenny P., Seitz, F., and O'Shaughnessy, D., "The use of minimum durations and energy contours for phonemes to improve large vocabulary isolated word recognition," submitted to IEEE Trans. on Acoustics, Speech, and Signal Processing, 1990.

15. Harrington, J., Watson, G., and Cooper, M., "Word boundary detection in broad class phoneme strings," Computer, Speech and Language, **3** (4), 367–382, 1989.

16. Jelinek, F., "The development of an experimental discrete dictation recognizer," Proc. IEEE, **73** (11), 1616–1624, 1985.

17. Kenny, P., Lennig, M., and Mermelstein, P., "Speaker adaptation in a large-vocabulary HMM recognizer," letter to the editor, IEEE Trans. Pattern Analysis and Machine Intelligence, August 1990.

18. Kenny, P., Lennig, M., and Mermelstein, P., "A linear predictive HMM for vector-valued observations with applications to speech recognition" IEEE Trans. on Acoustics, Speech, and Signal Processing, **38** (3), 220–225, 1990.

19. Lee, K.-F., Hon, H.-W., "Large-vocabulary speaker-independent continuous speech recognition using HMM," Proc. 1988 IEEE International Conference on Acoustics, Speech and Signal Processing, 123–126.

20. Merialdo, B., "Speech Recognition using very large size dictionary," Proc. 1987 IEEE International Conference on Acoustics, Speech and Signal Processing, 364–367.

21. Nilsson, N., *Principles of artificial intelligence*, Tioga Publishing Company, 1982.

22. O'Shaughnessy, D., "Using syntactic information to improve large-vocabulary word recognition," Proc. 1989 IEEE International Conference on Acoustics, Speech and Signal Processing, 44S13.6.

23. Poritz, A., "Hidden Markov models: a guided tour," Proc. 1988 IEEE International Conference on Acoustics, Speech and Signal Processing, 7–13.

24. Schwartz, R.M., Chow Y.L., Roucos S., Krasner M., and Makhoul J., "Improved hidden Markov modeling of phonemes for continuous speech recognition," Proc. 1984 IEEE International Conference on Acoustics, Speech, and Signal Processing, 35.6.1–35.6.4.

25. Seitz, F., Gupta, V., Lennig, M., Kenny, P., Deng, L., and Mermelstein, P., "A dictionary for a very large vocabulary word recognition system," Computer, Speech and Language, **4**, 193–202, 1990.

26. Seitz, F., Gupta, V., Lennig, M., Deng, L., Kenny, P.,and Mermelstein, P., "Phonological rules and representations in a phoneme-based very large vocabulary word recognition system," J. Acoust. Soc. Am. **87** (S1), S108, 1990.