Reconnaissance de la parole pour la dictée de documents

Continuous speech dictation

par J.-L. Gauvain, L. Lamel, M. Adda-Decker

LIMSI-CNRS
91403 Orsay Cedex

Résumé

Ceci est le résumé

Mots clés : reconnaissance de la parole

Abstract

One of our major research activities at LIMSI is multilingual, speaker-independent, large vocabulary speech dictation. The multilingual aspect of this work is of particular importance in Europe, where each country has it's own national language. Speaker-independence and large vocabulary are characteristics necessary to envision real world applications of such technology. The recognizer makes use of phone-based continuous density HMM for acoustic modeling and n-gram statistics estimated on newspaper texts for language modeling. The system has been evaluated on two dictation tasks developed with read, newspaper-based corpora, the ARPA Wall Street Journal corpus of American English and the BREF Le Monde corpus of French. Experimental results under closely matched conditions are reported. For both languages an average word accuracy of 95% is obtained for a 5k vocabulary test. For a 20,000 word lexicon with an unrestricted vocabulary test the word error for WSJ is 10% and for BREF is 16%.

Key words : *speaker-independent,speech recognition*

1. Introduction

Speech recognition research at LIMSI aims to develop recognizers that are task-, speaker-, and vocabulary-independent so as to be easily adapted to a variety of applications. The applicability of speech recognition techniques used for one language to other languages is of particular importance in Europe. The multilingual aspects are in part carried out in the context of the LRE SQALE (Speech recognizer Quality Assessment for Linguistic Engineering) project, which is aimed at assessing language dependent issues in multilingual recognizer evaluation. In this project, the same system will be evaluated on comparable tasks in different languages (English, French and German) to determine cross-lingual differences, and different recognizers will be compared on the same language to compare advantages of different recognition strategies.

In this paper some of the primary issues in large vocabulary, speaker-independent, continuous speech recognition for dictation are addressed. These issues include language modeling, acoustic modeling , lexical representation, and decoding. Acoustic modeling makes use of continuous density HMM with Gaussian mixture of context-dependent phone models. For language modeling n-gram statistics are estimated on text material. To deal with phonological variability alternate pronunciations are included in the lexicon, and optional phonological rules are applied during training and recognition. The decoder uses a time-synchronous graph-search strategy[20] for a first pass with a bigram back-off language model (LM)[13]. A trigram LM is used in a second acoustic decoding pass which incorporates the word graph generated in the first pass[10]. Experimental results are reported on the ARPA *Wall Street Journal (WSJ)*[24] and BREF[6, 14] corpora, using for both corpora over 37k utterances for acoustic training and more than 37M words of newspaper text for language model training. It has been shown[9] that for both corpora increasing the amount of

training utterances by an order of magnitude reduces the word error by about 30%. The use of a trigram LM in a second pass also gives an error reduction of 20% to 30%. The combined error reduction is on the order of 50%.

2. Définition du problème et fondements

In speech dictation we are principally concerned with the problem of transcribing the speech signal as a sequence of words. Today's most performant systems are for the most part based on a statistical modelisation of the talker. From this point of view, message generation is represented by a language model which provides estimates of $\Pr(w)$ for all word strings w, and the acoustic channel encoding the message w in the signal x is represented by a probability density function $f(x|w)$. The speech decoding problem consists then of maximizing the a posteriori probability of w, or equivalently, maximizing the product $\Pr(w)f(x|w)$.

The principles on which these systems are based have been known for many years now, and include the application of information theory to speech recognition[1, 12], the use of a spectral representation of the speech signal [4, 5], the use of dynamic programming for decoding[28, 29], and the use of context-dependent phone models[26, 3, 18]. Despite the fact that some these techniques were proposed well over a decade ago, considerable progress has been made in recent years that makes speaker-independent, continuous speech dictation feasible for vocabularies of at least 20,000 words. This progress has been substantially aided by the availability of large speech and text corpora and by significant advances made in micro-electronics which has enabled the development of more complex models and algorithms.

The same modeling techniques can be adapted to other related applications, such as speech understanding or spoken language systems or in the identification of what we can refer to as "non-linguistic" speech features[17]. These feature-specific models may also be directly used to more accurately model the speech signal thus in consequence improving the performance of the speech recognizer.

3. Modélisation du langage

Language modeling entails incorporating constraints on the allowable sequences of words which form a sentence. Statistical n-gram models attempt to capture the syntactic and semantic constraints by estimating the frequencies of sequences of n words. A backoff mechanism[13] is used to smooth the estimates of the probabilities of rare n-grams by relying on a lower order n-gram when there is insufficient training data, and to provide a means of modeling unobserved n-grams. Another advantage of the backoff mechanism is that LM size can be arbitrarily reduced by relying more on the backoff, by increasing the minimum number of required n-gram observations needed to include the n-gram. This property can be used in the first bigram decoding pass to reduce computational requirements. The LM training data consists of 37M words of the *WSJ* and 38M words of *Le Monde*. In order to be able to construct LMs for BREF, it was necessary to normalize the text material of *Le Monde* newspaper[8], which entailed a pre-treatment rather different from that used to normalize the *WSJ* texts[24].

Table ?? compares some characteristics of the *WSJ* and *Le Monde* text corpora. In the same size training texts, there are almost 60% more distinct words for *Le Monde* than for *WSJ* without taking case into account. If case is kept when distinctive (the numbers in parentheses), there are 280k words in the *Le Monde* training material. As a consequence, the lexical coverage for a given size lexicon is smaller for *Le Monde* than for *WSJ*. For example, the 20k *WSJ* lexicon accounts for 97.5% of word occurrences, but the 20k BREF lexicon only covers 94.9% of word occurrences in the training texts. For lexicons in the range of 5k to 40k words, the number of words must be doubled for *Le Monde* in order to obtain the same word coverage as for *WSJ*.

The lexical ambiguity is also higher for French than for English. The homophone rate (the number of words which have a homophone divided by the total number of words) in the 20k BREF lexicon is 57% compared to 9% in 20k-open *WSJ* lexicon. This effect is even greater if the word frequencies are taken into account. Given a perfect phonemic transcription, 23% of words in the *WSJ* training texts is ambiguous, whereas 75% of the words in the *Le Monde* training texts have an ambiguous phonemic transcription. Not only does one phonemic form correspond to different orthographic forms, there can also be a relatively large number of possible pronunciations for a given word. In French, the alternate pronunciations arise mainly from optional word-final phones, due to liaison and optional word-final consonant cluster reduction. There are also a larger number of frequent, monophone words for *Le Monde* than for *WSJ*, accounting for about 17% and 3% of all word occurrences in the respective training texts.

4. Modélisation acoustico-phonétique

The recognizer makes use of continuous density HMM (CDHMM) with Gaussian mixture for acoustic modeling. The main advantage continuous density modeling offers over discrete or semi-continuous (or tied-mixture) observation density modeling is that the number of parameters used to model an HMM observation distribution can easily be adapted to the amount of available training data associated to this state. As a consequence, high precision modeling can be achieved for highly frequented states without the explicit need of smoothing techniques for the densities of less frequented states. Discrete and semi-continuous modeling use a fixed number of parameters to represent a given observation density and therefore cannot achieve high precision without the use of smoothing techniques. This problem can be alleviated by tying some states of the Markov models. However, since this requires careful design and some a priori assumptions, these techniques are primarily of interest when the training data is limited and cannot easily be increased.

A 48-component feature vector is computed every 10 ms. This feature vector consists of 16 Bark-frequency scale cepstrum coefficients computed on the 8kHz bandwidth and their first and second order derivatives. The acoustic models are sets of context-dependent (CD), position independent phone models, which include both intra-word and cross-word contexts. The contexts are automatically selected based on their frequencies in the training data. The models include triphone models, right- and left-context phone models, and context-independent phone models. Each phone model is a left-to-right CDHMM with Gaussian mixture observation densities (typically 32 components). The covariance matrices of all the Gaussians are diagonal. Duration is modeled with a gamma distribution per phone model. The HMM and duration parameters are estimated separately and combined in the recognition process for the Viterbi search. Maximum a posteriori estimators are used for the HMM parameters[7] and moment estimators for the gamma distributions. Separate male and female models are used to more accurately model the speech data.

During system development phone recognition has been used to evaluate different acoustic model sets. It has been shown that improvements in phone accuracy are directly indicative of improvements in word accuracy when the same phone models are used for recognition[16]. Phone recognition provides the added benefit that the recognized phone string can be used to understand word recognition errors and problems in the lexical representation.

5. Représentation lexicale

The lexicons are represented phonemically,[1] using language-specific sets of phonemes. Alternate pronunciations are provided for about 10% of the words.[2] A pronunciation graph is generated for each word from the baseform transcription to which word internal phonological rules are optionally applied during training and recognition to account for some of the phonological variations observed in fluent speech.

Word boundary phonological rules are applied in building the phone graph used by the recognizer so as to allow for some of the phonological variations observed in fluent speech[15]. The principle behind the phonological rules is to modify the phone network to take into account such variations. These rules are optionally applied during training and recognition. Using phonological rules during training results in better acoustic models, as they are less "polluted" by wrong transcriptions. Their use during recognition reduces the number of mismatches. For English, only well known phonological rules, such as glide insertion, stop deletion, homorganic stop insertion, palatalization, and voicing assimilation have been incorporated in the system. The same mechanism has been used to handle liaisons, mute-e, and final consonant cluster reduction for French.

6. Statégie de décodage

One of the most important problems in implementing the decoder of a large vocabulary speech recognizer is the design of an efficient search algorithm to deal with the huge search space, especially when using language models with a longer span than two successive words, such as trigrams. The most commonly used approach for small and medium vocabulary sizes is the one-pass frame-synchronous beam search [20] which uses a dynamic programming procedure. This basic strategy has been recently extended by adding other features such as "fast match"[11, 2], N-best rescoring[27], progressive search[19] and one-pass dynamic network decoding[21]. The two-pass approach used in our system is based on the idea of progressive search where the

[1] The lexicons were all developed at LIMSI. For French, the base pronunciations were obtained using text-to-phoneme rules[25] and extended to annotate potential liaisons and pronunciation variants.

[2] This does not count word final optional phonemes marking possible liaisons for French. Including these raises the number of entries with multiple transcriptions to almost 40%.

information between levels is transmitted via word graphs. Prior to word recognition, sex identification is performed for each sentence using phone-based ergodic HMMs[17]. The word recognizer is then run with a bigram LM using the acoustic model set corresponding to the identified sex.

The first pass of the decoder uses a bigram-backoff LM with a tree organization of the lexicon for the backoff component. This one-pass frame-synchronous beam search, which includes intra- and inter-word CD phone models, intra- and inter-word phonological rules, phone duration models, and gender-dependent models, generates a list of word hypotheses resulting in a word lattice. Two considerations need to be taken into account at this level. The first is whether or not the dynamic programming procedure used in the first pass, which guarantees the optimality of the search for the bigram, generates an "optimal" lattice to be used with a trigram LM. For example, any given word in the lattice will have many possible ending points, but only a few starting points. This problem is in fact less severe than expected since the time information is not critical to generate an "optimal" word graph from the lattice, i.e. the multiple word endings provide enough flexibility to compensate for single word beginnings. The second consideration is that the lattice generated in this way cannot be too large or there is no interest in a two pass approach. To solve this second problem, two pruning thresholds are used during the first pass, a beam search pruning threshold which is kept to a level insuring almost no search errors (from the bigram point of view) and a word lattice pruning threshold used to control the lattice size.

The following steps give the key elements behind the procedure used to generate the word graph from the word lattice.[3] First, a word graph is generated from the lattice by merging three consecutive frames (i.e. the minimum duration for a word in our system). Then, "similar" graph nodes are merged with the goal of reducing the overall graph size and generalizing the word lattice. This step is reiterated until no further reductions are possible. Finally, based on the trigram backoff language model a trigram word graph is then generated by duplicating the nodes having multiple language model contexts. Bigram backoff nodes are created when possible to limit the graph expansion.

It should be noted that this decoding strategy based on two forward passes can in fact be implemented in a single forward pass using one or two processors. We are using a two pass solution because it is conceptually simpler, and also due to memory constraints.

7. Résultats expérimentaux

The recognizer was evaluated under closely matched conditions for American English and for French, with vocabularies of 5k and 20k words. For French the 20k test included both open and closed vocabulary data. The training data (see Table ??) include about 38k sentences for each language. The standard *WSJ0/WSJ1* SI284 training material containing 37,518 sentences from 284 speakers was used for English. For French, the BREF training data contains 38,550 sentences from 80 speakers.

The *WSJ* system was evaluated in the Nov92 ARPA evaluation test[22] for the 5k-closed vocabulary and in the Nov93 ARPA evaluation test[23] for the 5k and 20k/64k hubs.[4] The word errors using 3306 CD models are given in Table ??. With a bigram LM, word errors of 4.8% and 6.8% are obtained respectively on the Nov92 and Nov93 5k test data. The trigram second pass reduces the word error by 35% on the Nov92 test data and by 22% on the Nov93 test data. On the 20k-open Nov92 and Nov93 test data the word errors with a bigram LM are 11.0% and 15.2%. In this open-vocabulary test data there are slightly over 2% out-of-vocabulary (OOV) words. Using the trigram LM reduces the error rate by about 20%.

Recognition results for BREF are given in Table ?? for 1747 CD models with bigram and trigram LMs. The word error on the 5k test data is 9.0%. The use of a trigram LM gives an error reduction of 39% to 5.5%. The word errors on the 20k and 20k-open test data are 12.9% and 19.5% respectively with the bigram LM. The use of the trigram LM reduces the word error by an additional 29% for the closed vocabulary test data, but only 16% on the open vocabulary test data. This difference can be attributed to the 3.9% of the words which are OOV and occur in 72 of the 200 test sentences. There is almost a 50% increase in word error, including a three-fold increase in word insertions compared with the closed vocabulary test. Thus apparently the OOV words are not simply replaced by another word, but are more often replaced by a sequence of words.

[3] In our implementation, a word lattice differs from a word graph only because it includes word endpoint information.

[4] The 20k open test for *WSJ* is also referred to as a 64k test since all of the words in these sentences occur in the 63,495 most frequent words in the normalized *WSJ* text material[24].

8. Conclusion

In this paper we have addressed some of the major issues in large vocabulary, speaker-independent, continuous speech dictation. These include acoustic modeling, language modeling, modeling phonological variations, and decoding. Experimental results have been presented for English and French using 5k and 20k vocabularies. Word accuracies on the order of 95% have been obtained for 5k vocabularies for both languages. With 20k lexicons and an open vocabulary test the word error is on the order of 10% for *WSJ* and 16% for BREF. This difference in word error can be largely attributed to the larger number of out-of-vocabulary words in French, an effect of the lower word coverage for the lexicon. Performance levels on this general news dictation task are sufficient to envision commercial application of this technology on simpler tasks in particular domain areas, such as dictation of medical, legal, police reports, insurance claims and contracts and other professional limited domain documents.

8. Remerciements

The authors wish to thank Gilles Adda for preprocessing and normalizing the *Le Monde* text materials.

BIBLIOGRAPHIE

[1] L. Bahl et al., "Preliminary results on the performance of a system for the automatic recognition of continuous speech," *ICASSP-76.*

[2] L.R. Bahl et al, "A Fast Match for Continuous Speech Recognition Using Allophonic Models," *ICASSP-92.*

[3] Y.L. Chow et al., "The Role of Word-Dependent Coarticulatory Effects in a Phoneme-Based Speech Recognition System", *ICASSP-86.*

[4] J. Dreyfus-Graf, "Sonograph and SounSpeaker-InSpeaker-Ind Mechanics," *JASA*, **22**, 1949.

[5] H. Dudley, S. Balashek, "Automatic Recognition of Phonetic Patterns in Speech," *JASA*, **30**, 1958.

[6] J.L. Gauvain, L.F. Lamel, M. Eskénazi, "Design considerations & text selection for BREF, a large French read-speech corpus," *ICSLP-90.*

[7] J.L. Gauvain, C.H. Lee, "Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities," *Speech Communication*, **11**(2-3), 1992.

[8] J.L. Gauvain et al., "Speaker-InSpeaker-Independent Continuous Speech Dictation," *Eurospeech-93.*

[9] J.L. Gauvain et al., "The LIMSI Continuous Speech Dictation System," *ARPA HLT Workshop*, 1994.

[10] J.L. Gauvain et al., "The LIMSI Continuous Speech Dictation System : Evaluation on the ARPA Wall Street Journal Task," *ICASSP-94.*

[11] L. Gillick, R. Roth, "A Rapid Match Algorithm for Continuous Speech Recognition," *DARPA Sp&NL Workshop*, 1990.

[12] F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proc. of the IEEE*, **64**(4), april 1976.

[13] S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. ASSP*, **35**(3), 1987.

[14] L.F. Lamel, J.L. Gauvain, M. Eskénazi, "BREF, a Large Vocabulary Spoken Corpus for French," *Eurospeech-91.*

[15] L. Lamel, J.L. Gauvain, "Continuous Speech Recognition at LIMSI," Final review *DARPA ANNT Speech Prog.*, Sep. 1992.

[16] L. Lamel, J.L.Gauvain,"High Performance Speaker-Independent Phone Recognition Using CDHMM," *Eurospeech-93.*

[17] L. Lamel, J.L. Gauvain, "Identifying Non-Linguistic Speech Features," *Eurospeech-93.*

[18] K.F. Lee, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition : The SPHINX System," *Ph.D. Thesis, CMU*, 1988.

[19] H. Murveit et al, "Large-Vocabulary Dictation using SRI's Decipher Speech Recognition System : Progressive Search Techniques," *ICASSP-93.*

[20] H. Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition," *IEEE Trans. ASSP*, **32**(2), pp. 263-271, April 1984.

[21] J.J. Odell et al., "A One Pass Decoder Design for Large Vocabulary Recognition," *ARPA HLT Workshop*, 1994.

[22] D.S. Pallett et al., "Benchmark Tests for the DARPA Spoken Language Program," *ARPA HLT Workshop*, 1993.

[23] D.S. Pallett et al., "1993 Benchmark Tests for the ARPA Spoken Language Program," *ARPA HLT Workshop*, 1994.

[24] D.B. Paul, J.M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *ICSLP-92.*

[25] B. Prouts,"Contribution à la synthèse de la parole à partir du texte : Transcription graphème-phonème en temps réel sur microprocesseur", Thèse de docteur-ingénieur, Université Paris XI, Nov. 1980.

[26] R. Schwartz et al, "Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition," *ICASSP-84.*

[27] R. Schwartz et al.,"New uses for N-Best Sentence Hypothesis, within the BYBLOS Speech Recognition System," *ICASSP-92.*

[28] T.K. Vintsyuk, "Speech discrimination by dynamic programming," *Kibnernetika*, **4**, 1968.

[29] T.K. Vintsyuk, "Elements-wise recognition of continuous speech composed of words from a specified dictionary," *Kibernetika*, **7**, 1971.