Large Vocabulary Speech Recognition in English and French

J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{gauvain,lamel,adda,madda}@limsi.fr

In this paper we report efforts at LIMSI in speaker independent large vocabulary speech recognition in French and in English. The recognizer makes use of continuous density HMM (CDHMM) with Gaussian mixture for acoustic modeling and n-gram statistics estimated on text material for language modeling. Acoustic modeling uses cepstrum-based features, context-dependent phone models (intra and inter-word), phone duration models, and sex-dependent models.

The recognizer uses a time-synchronous graph-search strategy as opposed to some recently developed multi-level approaches, i.e. multiple pass strategies[13, 1, 9]. We show that the time-synchronous approach is still viable with vocabularies of up to 20K words when used with bigram backoff language models. This one level implementation includes intra- and inter-word CD phone models, intra- and inter-word phonological rules, phone duration models, and bigram-backoff language model[6, 3]. The backoff mechanism has been efficiently implemented using a lexicon tree. Using this approach the search space can be arbitrarily reduced by relying more on the backoff. The HMM-based word recognizer graph is built by putting together word models according to the grammar in one large HMM. Each word model is obtained by concatenation of the phone models, according to its phone transcription as found in the lexicon.

We have demonstrated that phonological rules are helpful even when using context-dependent phone models. The principle behind the phonological rules is to modify the phone network to take into account phonological variations. The rules are applied during both training and recognition and are always optional. Using optional 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. The mechanism for phonological rules allows the potential for generalization and extension.

Much of the development has been carried out using the Resource Management Corpus and also by performing phone recognition on Wall Street Journal instead of word recognition in order to reduce the computational requirements and speed up the development process. We have found that improvements in phone accuracy are directly indicative to improvements in word accuracy when the same phone models are used for recognition[3, 7]. This has allowed us to evaluate many alternatives for the front-end and the acoustic models.

Evaluation of the system has been based on read speech corpora: DARPA Wall Street Journal-based CSR corpus and the BREF corpus containing recordings of texts from the French newspaper *Le Monde*. For both corpora experiments were carried out with up to 20K word lexicons. Moving from the Resource Management task to Wall Street Journal required rewritting the recognition engine in order to deal with the vast difference in the size of the recognition graph. For RM, with the standard word pair grammar there were on the order of 60,000 word connections. In contrast, the 20,000 word Wall Street Journal Task with a word bigram may have several million word connections. In the remainder of this paper we present experimental results on the WSJ and BREF corpora with similar size lexicons and task complexities and point out language-specific properties and problems.

EXPERIMENTS WITH WSJ

The standard WSJ0 SI-84 training data include 7240 sentences from 84 speakers. The language model is a bigram-backoff estimated on the 33 million word standardized WSJ text provided by Lincoln Labs[10]. The lexicon is represented using a set of 46 phones. The pronunciations were obtained from various existing lexicons (TIMIT, Pocket and Moby), missing forms were generated by rule when possible, or added by hand. Some of the missing proper names were transcribed by the ORATOR system of Bellcore.

This system was evaluated in the Nov92 DARPA evaluation test for the 5k-closed vocabulary using the standard bigram language models[10]. The official reported results for NVP are given in the first line of Table 1 using 493 CD models, without the second derivative of the cepstral coefficients. Increasing the number of CD models and the number of features, reduced the error rate by about 20% over the system used for the Nov92 evaluation. Results are also given in Table 1 for the Nov92 NVP 64k test data using both open and closed 20k vocabularies. (The 20k closed vocabulary includes all the words in the test data whereas the 20k open vocabulary contains only the 20k most common words in the WSJ texts[10]). It can be seen that the error rate is doubled when the vocabulary size goes from 5k to 20k, whereas the test perplexity goes from 111 to 244. The higher error rate with the 20k+ open lexicon can be contributed to the out-of-

WSJ - Conditions	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Err.*
493m, 32f, 5k*	91.8	6.9	1.3	1.5	9.7
884m, 48f, 5k	94.1	5.2	0.7	1.0	6.9
884m, 48f, 20k	88.3	10.1	1.5	2.0	13.6
884m, 48f, 20k+	86.8	11.7	1.5	2.7	15.9

Table 1: Word recognition results on the WSJ0 corpus with a probabilistic grammar (2-grams) estimated on WSJ text data. 5k: 5000 word lexicon, 20k: 20,000 word lexicon, 20k+: 20,000 word lexicon with open test, *official DARPA NOV92 evaluation results.

BREF - Conditions	*Corr.*	*Subs.*	*Del.*	*Ins.*	*Err.*
428m, 5k	87.1	10.3	2.6	1.7	14.5
428m, 20k	84.6	12.8	2.6	2.9	18.3
428m, 5k-H	90.7	7.2	2.6	1.7	11.5
428m, 20k-H	89.4	8.0	2.6	3.0	13.5

Table 2: Word recognition results on the BREF80 corpus with a probabilistic grammar (2-grams) estimated on *Le Monde* text data. 5k: 5000 word lexicon, 20k: 20,000 word lexicon, 5k-H: 5k word lexicon with homophone errors not counted.

vocabulary words, which account for almost 2% of the words in the test sentences.

One problem using the bigram language model is that the number of connections is very large. We investigated the effects of reducing the size of the bigram model by relying more on the backoff. Using a count threshold of 4 occurrences, reduces the bigram size by 53% and gives a word error of 7.2% on the 5k test. This is only a slight increase in the error compared to the 6.9% obtained with a threshold of 1 (baseline bigram[10]).

EXPERIMENTS WITH BREF

BREF is a large read-speech corpus, containing over 100 hours of speech material, from 120 speakers (55m/65f)[8]. The speech data used in these experiments come from the BREF80 sub-corpus (2 CDs). 2770 sentences from 57 speakers were used for training. Phonetic transcriptions of the training data were automatically derived and manually verified[2]. A bigram-backoff language model was estimated on about 4 million words of normalized text material from *Le Monde*. The base lexicon, represented with 35 phones, was obtained using text-to-phoneme rules, and was extended to annotate potential liaisons and pronunciation variants. As for the WSJ task, two vocabularies have been used for the recognition experiments, corresponding to the 5k and 20k most common words in the *Le Monde* texts. The test data consist of 100 sentences for each vocabulary size, with perplexities of 122 for the 5k sentences and 205 for the 20k sentences.

Word recognition results using 428 CD models and the bigram-backoff language model estimated on the normalized text material from *Le Monde* are shown in Table 2. The word error is 14.5% for the 5k lexicon and 18.3% for the 20k lexicon. We can see that the word recognition error for French is substantially higher than for English with similar tasks and test perplexities. Part of this is due to the larger number of homophones for French, as well as larger homophone classes[4, 5]. The last two entries give the results when the recognizer output is scored without counting single word homophone errors. Many more complex homophones still remain such as the *multiple word* "leur coût" (sing.) and "leurs coûts" (pl.), or *multiword* homophones such as "a mis" and "amis". The difference in the results scored with and without homophones points out the need for better language modeling.

DISCUSSION AND SUMMARY

Some differences in the characteristics of French and English have become apparent as a result of these experiments. The lexical coverage for French (BREF: 5k (86%), 20k (95%)) is significantly smaller than for the same size lexicons for English (WSJ: 5k (92%), 20k (98%)). The lexicon size for BREF must be doubled in order to obtain the same coverage as for WSJ. As pointed out earlier, homophones are much more common in French - in the BREF training texts one out of two words are homophones, compared to one in five in the WSJ training texts. The largest homophone class in the WSJ lexicon has 4 entries: *B., Bea, bee, and be*. In the BREF lexicon there are 3 homophone classes each having 7 orthographic words, as in *100, cent, cents, san, sang, sans, sent*.

In French we must also deal with the problem of liaison, which does not occur in English. In part due to liaison, there can be a relatively large number of pronunciations for a given word. For example, the word "autres" has the following transcriptions: /ot/, /otrx/, /otr/, /otrxz/, each of which is possible, but not equally likely, depending on the speaker, the dialect, the neighboring phones and words, and sometimes on the semantics.

It is interesting to note that while French has higher word error rates than English for comparable size tasks, phone recognition in French is better than in English. The phone accuracy on the same test data for BREF using the same 428 CD phone sets is 78.7% (phone px=16.1), compared to 74.1% for WSJ0 with 884 CD phone models (72.4% 488 models) (phone px=17.5).[7]. Therefore we conclude that the difference in performance must be due to the lexical confusability for French.

We have observed that for both French and English, a large number of recognition errors involve short words of one or two phonemes. While there are relatively few of these words, they are very frequent, accounting for about 50% of all words occurrences in French and 30% of all word occurrences in English. In particular there are more monophone words in French, and they are much more frequent than in English. Almost 20% of the words in the BREF training texts are single phone words, compared to only 3% in the WSJ training texts. Since it is relatively easy to insert and delete monophone words, it is expected that French should have a higher word error rate than English. The large number of monophone words and the need to deal with liaison result in a

search space for French that is twice as large as is needed for the same size vocabulary in English. The difference in search space size for open vocabulary systems is even larger if we want to have comparable lexical coverage for the two languages.

The evaluations and extensive error analysis have pointed out different problems that must be dealt with in the two languages. One of our goals is to determine how much effort and of what type is involved in porting to a new language or a new task.

REFERENCES

[1] F. Alleva, X. Huang, M.-Y. Hwang, "An Improved Search Algorithm Using Incremental Knowledge for Continuous Speech Recognition," *ICASSP-93*.

[2] J.L. Gauvain, L.F. Lamel, "Speaker-Independent Phone Recognition Using BREF," *DARPA Sp. & Nat. Lang. Workshop*, Feb. 1992.

[3] J.L. Gauvain, L.F. Lamel, G. Adda, "LIMSI Nov92 WSJ Evaluation," presented at the *DARPA Spoken Language Systems Technology Workshop*, Cambridge, MA, Jan. 1993.

[4] J.L. Gauvain et al, "Speaker-Independent Continuous Speech Dictation," *EUROSPEECH-93.*

[5] J.L. Gauvain et al, "Speech-to-Text Conversion in French," to appear in *Int. J. Pat. Rec. & A.I.*, 1994.

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

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

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

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

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

[11] D.S. Pallett, J.G. Fiscus, "Resource Management Corpus - Continuous Speech Recognition - September 1992 Test Set Benchmark Test Results," Finalf review *DARPA ANNT Speech Program*, Sep. 1992.

[12] D.S. Pallett et al, "Benchmark Tests for the DARPA Spoken Language Program," *ARPA Workshop on Human Language Technology*, Mar. 1993.

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