Continuous Speech Recognition at LIMSI

Lori F. Lamel and Jean-Luc Gauvain

LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{lamel,gauvain}@limsi.fr

ABSTRACT

This paper presents some of the recent research on speaker-independent continuous speech recognition at LIMSI including efforts in phone and word recognition for both French and English. Evaluation of an HMM-based phone recognizer on a subset of the BREF corpus, gives a phone accuracy of 67.1% with 35 context-independent phone models and 74.2% with 428 context-dependent phone models. The word accuracy is 88% for a 1139 word lexicon and 86% for a 2716 word lexicon, using a word pair grammar with respective perplexities of 101 and 160. Phone recognition is also shown to be effective for language, sex, and speaker identification.

The second part of the paper describes the recognizer used for the September-92 Resource Management evaluation test. The HMM-based word recognizer is built by concatenation of the phone models for each word, where each phone model is a 3-state left-to-right HMM with Gaussian mixture observation densities. Separate male and female models are run in parallel. The lexicon is represented with a reduced set of 36 phones so as to permit additional sharing of contexts. Intra- and inter-word phonological rules are optionally applied during training and recognition. These rules attempt to account for some of the phonological variations observed in fluent speech. The speaker-independent word accuracy on the Sep92 test data was 95.6%. On the previous test materials which were used for development, the word accuracies are: 96.7% (Jun88), 97.5% (Feb89), 96.7% (Oct89) and 97.4% (Feb91).

INTRODUCTION

LIMSI, Laboratory of Computer Science for Mechanical and Engineering Science is part of the CNRS, a national center for scientific research. This is a French government organization, which employs over 27,000 people, two-thirds of whom are researchers, and the remaining third are technical and support staff. LIMSI has two departments: Mechanics/Energetics (40 researchers) and Man-Machine Communication (70 researchers). The latter is further divided into groups for Speech Communication, Language and Cognition, and Non-Verbal Communication.

The activities of the Speech Communication Group (14 permanent CNRS researchers) include speech recognition, speech synthesis, speaker verification and identification, and dialog. The continuous speech recognition (CSR) efforts are directed at spoken language systems and at speech-to-text decoding. For both applications, a speaker-independent (SI), vocabulary-independent (VI), phone recognizer is being developed, so as to be easily adaptable to various tasks. A dialog project oriented toward Air-Traffic Controller training[17], in collaboration with the National Center for Air-Traffic Control (CENA), is undergoing on-site performance evaluations. Without this system, the student training sessions require a human instructor who plays the roles of the pilots. The goal is to replace the instructor by a spoken dialog system so as to increase the availability of the system for training. The dialog system is built around the *Amadeus* speech recognizer and an associated synthesis module. Another dialog project has recently been initiated in collaboration with MIT to bring up a French ATIS system. In speech-to-text, the research focuses on two tasks: BREF[8, 14, 7] and Wall Street Journal (WSJ)[18].

LIMSI has been involved in the development of real-time recognizers for over 10 years, including the first French single-board isolated-word speech recognizer, *Moise*, and the first single-board connected-word recognizer, *Mozart* [6]. The *Amadeus* speaker-dependent recognizer was developed around a custom VLSI search processor (μPCD)[19] that was designed at LIMSI, in collaboration with the Bull and the Vecsys companies. This dedicated processor is fully programmable and can support isolated-word (5000 words) and connected-word recognition (300 words) algorithms using DTW or HMM approaches in real-time. The vocabulary size can be extended by multiplying the number of such processors; a single IBM PC board can support up to 16 processors.

THE BREF CORPUS

The research in the speech-to-text has been primarily focused on French, using the BREF corpus [8, 14]. Recently, work has also been started on the DARPA continuous speech recognition (CSR) task, using the Wall Street Journal corpus. The immediate goal is to work with read speech material from a large number of speakers, so as to be able to build base acoustic models which can be augmented and adapted to specific speakers or tasks. This approach also allows many aspects of language modeling to be addressed under more "semi-controlled conditions," than those found in spontaneous dictation. Additionally, it is much easier to collect read-text material than spontaneous dictations.

BREF is a large read-speech corpus, containing over 100 hours of speech material, from 120 speakers (55m/65f)[14]. The text materials were selected verbatim from the French newspaper *Le Monde*, so as to provide a large vocabulary (over 20,000 words) and a wide range of phonetic environments[8]. Containing 1115 distinct diphones and over 17,500 triphones, BREF can be used to train VI phonetic models.

In these experiments approximately 4 hours and 20 minutes of speech material are used for training. This represents 2770 sentences from 57 speakers (28m/29f). The test data consist of 109 sentences from 19 speakers (10m/9f). The test text material is distinct from the training texts, and the test speech data contain 7635 phone segments. Phonemic transcriptions of these utterances were automatically generated and manually verified[7].

PHONE RECOGNITION WITH BREF

Evaluating phonetic recognition is important for several reasons. Primarily, the demands of VI, SI, CSR require an approach based on phone-like units. The better these phone models (or acoustic models) are, the better the performance of the entire system will be. Only considering word recognition performance, particularly when word-based grammars are used, can mask problems that stem from the acoustic level. Phone recognition is also useful in determining pronunciation errors in the lexicon and identifying alternate pronunciations that need to be included.

The phone recognizer uses a set of phone models, where each phone model is a 3-state left-to-right continuous density hidden Markov model (CDHMM) with Gaussian mixture observation densities. The covariance matrices of all the Gaussians components are diagonal. Silence is treated as a phone, but is modeled with a 1-state HMM. The 16 kHz speech was downsampled by 2 and a 26-dimensional feature vector was computed every 10 ms. The feature vector is composed of 13 cepstrum coefficients and 13 differential cepstrum coefficients. Duration is modeled with a gamma distribution per phone model. As proposed by Rabiner et al.[20], the HMM and duration parameters are estimated separately and combined in the recognition process for the Viterbi search. Maximum likelihood estimators are used for the HMM parameters[11] and moment estimators for the gamma distributions.

For context-independent (CI) models, the overall Markov chain is simply obtained by allowing all possible connections between the 35 phone HMMs (i.e. 1225 connections). For the transition probabilities either constant (1/35), 1-gram, or 2-gram probabilities are used. The resulting ergodic HMM has 103 states and about 170,000 parameters.

In the case of context-dependent (CD) models, the phone HMMs are connected through null states representing all the possible diphones. These null states, which do not emit any observation, are used to merge all the transitions corresponding to the same diphone, thus reducing the number of connections to a more manageable value (i.e., the fourth order (n^4) becomes a cubic form). With 428 CD models, the resulting HMM includes 1294 non-null states and has about 1,070,00 parameters.

Table 1 gives the phone accuracy using 35 CI phone models and 428 CD phone models with 16 mixture components. Silence segments were not included in the computation of the phone accuracy. It can be seen that the phone language model helps more for CI than the CD models. This is presumably because the CD models already incorporate some of the phonotactic information. With the phone bigram, the use of CD models reduces the errors by 22%. In these experiments using as many as 2100 CD models did not significantly reduce the error rate.

Model set	*0-gram*	*1-gram*	*2-gram*
35 CI	61.0	63.4	67.1
428 CD	71.2	71.9	74.2

Table 1: Phone accuracy with CI and CD models.

WORD RECOGNITION WITH BREF

In these word recognition experiments the same set of 428 CD phone models were used. An HMM is generated for each word by concatenating the phone models according to the phone transcriptions and the word models are put together to represent the entire lexicon with one large HMM. For the no-grammar (NG) case a phone tree is built from the lexicon in order to reduce the graph size. For the 10K lexicon the average number of phone nodes per word is reduced from 6.4 to 2.0 by using such a tree instead of a linear representation of each word, giving a reduction of 69% in the size of the graph. For the word-pair grammar (WPG), a phone graph is first built by linking the word phone transcriptions according the grammar, then, as for the no-grammar case, the phone graph is converted to a large HMM by replacing each phone node by the appropriate set of phone models and establishing the proper connections with the neighboring phones. In both cases, CD phone models are used for word juncture phones as well as for intra-word phones, without explicit representation of the word boundaries. Recognition is performed using the Viterbi decoding algorithm.

Vocabulary-independent word recognition experiments were run using four different lexicons. The 1K lexicon contains only the 1139 words found in the test sentences. The 3K lexicon contains all the words found in the earlier training sentences[7], and the test sentences, a total of 2716 words. The 5K (4863 words) and 10K (10,511 words) lexicons include all the words in the test data complemented with the most common words in the original text. Alternate pronunciations increase the number of phonemic forms in the lexicon by about 10%. The word recognition results are given in Table 2 with no grammar and with a word-pair grammar computed on the entire 4.2 million word text of *Le Monde*. For the no grammar condition, single word homophone confusions were not counted as errors. The use of the word-pair grammar reduces the perplexities to 101 for the 1K lexicon and 160 for the 3K lexicon, and reduces the error rate by almost 60%. In addition, the drop in performance observed by increasing the lexicon size is smaller than for the no grammar case, as is expected given that the perplexity is not proportional to the size of the lexicon.

No Grammar				*WP Grammar*	
1K	*3K*	*5K*	*10K*	*1K (101)*	*3K (160)*
72.7	66.5	62.0	59.7	87.9	86.1

Table 2: Word accuracy with 1K-10K lexicons.

The large number of homophones presents problems in phoneme-to-text conversion of French. The *Le Monde* text lexicon has a homophone rate of about 30%, compared to roughly 3% in the DARPA RM lexicon and under 2% for the DARPA TIMIT lexicon[3]. In French one must also deal with "liaison", "mute-e", and "apostrophe." Liaisons are links made between words, phonemes that are pronounced at the junctions between two words, but would not be pronounced at the end of the first word, or at the beginning of the second one, if the words were spoken in isolation. The pronunciation of mute-e is optional and dialect dependent, and poses problems similar to that of schwa-deletion in English. For apostrophe, the final vowel of certain words is deleted when the next word begins with a vowel. While in the written form, an apostrophe replaces the vowel, in the spoken form, there is no replacement.

IDENTIFICATION OF NON-LINGUISTIC SPEECH FEATURES

Phone recognition has also been found to be effective for identifying non-linguistic speech features, such as the sex of the speaker, the identity of the speaker, and the language spoken. In these studies CD models are used for sex identification and CI models are used for speaker and language identification.

Sex Identification

It is well known that the use of sex-dependent models gives improved performance over one set of speaker-independent models. However, this approach is costly in terms of computation for even medium-size tasks. A logical extension is to use first phonetic recognition to determine the speaker's sex, and then perform word recognition using the models of selected sex. This is the approach that we use in the WSJ system. Phone recognition using CD male and female models was performed, and the sex of the speaker was selected as the sex associated with the models that had the highest likelihood. No errors were observed in sex-identification for WSJ or for BREF data.

Speaker Identification

The same approach has also been applied to speaker identification. In this case a set of phone models were built for each speaker, by supervised adaptation of SI models[9]. The unknown speech was recognized by all of the speakers models in parallel. Experiments for English using the 462 training speakers in the TIMIT corpus[3] resulted in 99.6% correct identification using one sentence for indentification, and 100% identification if the likelihood over two sentences was used. A simple reduction in computation is gained by first determining the sex of the speaker by running in parallel SI male and female models. In experiments with this approach no cross-sex errors have ever occurred with the SI male/female models or with any of the SD models. Further reductions in the computation required during recognition can be obtained by speaker clustering.

Language Identification

Another application for phonetic recognition is language identification. The basic idea is to process in parallel the unknown incoming speech by different sets of phone models for each of the languages under consideration, and to choose the language associated with the model set providing the highest likelihood. Experiments have been performed using sets of SI CI phone models for French and for English[13]. For French the set of 35 SI CI models were used. For English a set of 52 SI CI phone models were trained on the training speakers in the TIMIT Corpus[3]. Using this approach and processing the entire utterance always gave 100% correct language identification. The identification accuracy as a function of the duration of the incoming speech is given in Table 3. The results are for increasing portions of speech taken from sentences spoken by 8 speakers (4m/4f) of each language. It was found that while with even as little as 400 ms of speech, the English sentences were always correctly identified, a minimum duration of one second was needed for perfect identification of French. This assymmetry may be because English has more phonemes than French, or that most of the French phonemes are also found in English. Additionally, the French phonemes not in English are acoustically not very different from allophones in English.

Duration	*400 ms*	*600 ms*	*800 ms*	*≥1000 ms*
French	72	75	87	100
English	100	100	100	100

Table 3: Language identification as a function of duration and language.

While the above results show this approach to reliably distinguish the two languages, there are differences in the corpora, which may have also influenced the results. In order to minimize these differences, the experiment was performed for French and English utterances spoken by a bilingual male Canadian speaker. The same BREF-based and TIMIT-based CI phone models were used, and the test sentences were taken from an Air Travel Information Services task, where the English and French versions were translations. The same trend was observed: the English sentences were always identified as English, and French sentences longer than 600 ms were also always correctly identified. Extensions of this work include identification of other European languages.

RM-SEP92 SYSTEM

In this section a detailed description of the recognizer used in the Sep92 DARPA evaluation test is given. The recognizer is basically the same as was used for the studies on BREF. Differences are essentially in the front end, the phone set, and the incorporation of phonological rules. The JUN88, FEB89, and OCT89 test sets were used as development data to evaluate various alternatives for the front end, the lexicon representation and phonological rules, and to estimate some parameter values such as the word insertion penalty.

System description

The main characteristics of the Sep92 system are:

Signal analysis A 48-component feature vector is computed every 10 ms. The feature vector consists of 16 Mel-frequency scale cepstrum coefficients and their first and second order differences.

Acoustic models There are about 2300 acoustic models of context-dependent phones. The contexts include both intra-word and cross-word contexts, but are position independent. Each phone model is a left-to-right CDHMM with an average of 10 gaussians per state. Duration is modeled with a gamma distribution per phone model. Separate male and female models are used.

Lexicon The lexicon is represented using a reduced set of 36 phones in order to better share contexts and to eliminate estimation problems due to infrequent phones. The lexicon has alternate pronunciations for about 10% of the words, and also allows some phones to be optional. For example, the word "MONTICELLO" has the pronunciations /mantxsElo/ and /mantxtSElo/, and the /t/ in "COUNTED" (/kawn{t}xd/) is optional. Intra- and inter-word phonological rules are optionally applied during training and recognition. These rules attempt to account for some of the phonological variations commonly observed in fluent speech, such as palatalization and glide insertion.

Decoding Decoding consists of Viterbi search, a one pass beam search. The male and female models are run in parallel, and the output with the highest likelihood is chosen. Again, in all of the development and evaluation test material, no cross-sex confusions ever occurred, i.e. never was a higher likelihood obtained using models of the other sex.

Front end

Experiments were run varying the analysis used to compute the cepstrum coefficients (LPC or Fourier analysis), and for two bandwidths (4kHz and 8kHz). The best performance was obtained with a 48-component feature vector consisting of 16 Bark-frequency scale cepstrum coefficients and their first and second order differences,

Diphthongs:			Syllabics:		
Y	→	ai	L	→	xl
O	→	ci	N	→	xn
W	→	aw	Contextual allophones:		
Infrequent phones:			|	→	x
U	→	∧	X	→	R
Z	→	z	Affricates:		
ϵ	→	n	C	→	tS

Table 4: Mapping of the eliminated phones.

computed on the 8kHz bandwidth. For each frame (30 ms window), a 15 channel Bark power spectrum is obtained by applying triangular windows to the DFT. The cepstrum coefficients are then computed using a cosinus transform [2]. On the development data, using an 8kHz bandwitdh instead of 4kHz reduced the error rate by 19% for a given set of SI CD phone models. For the 4kHz bandwidth, no significant difference was observed by using LPC or DFT based cepstra.

Why a reduced phone set?

Two different phone sets were evaluated. The first includes 47 phones (AT&T RM phone set[15]) and the second is a reduced set of 36 phones. The phone set was reduced primarily to eliminate infrequent phones for which there was insufficient training data, and to provide a means of better sharing contexts. In doing so, there is more data available to train the models, and the number of potential triphone contexts is reduced. This is a kind of parameter sharing.

The changes made to the 47 phone set are given in Table 4, using the one character MIT TIMIT symbol set[3]. Phones which were infrequent such as /Z/ and /U/ were eliminated, and replaced by another "close" phone. Certain phonemic distinctions are somewhat artibrary, such as whether the diphthongs should be represented as one vowel or a sequence of two vowels. While in the 47 phone set, diphthongs are represented distinctly, in the reduced set, the diphthongs /Y,O,W/ are represented by a sequence of phones. Similarly, allophonic distinctions such as the syllabics, the context-dependent difference between the two schwas (/x,|/), and the stress difference between /X,R/, are no longer made. Care was taken to ensure that these changes did not create any new homophones in the lexicon. Reducing the phone set gave an improvement of about 10% on the 3 development tests.

After varying the signal analysis and parameter vector, and reduction of the phone set, testing on unseen data from the FEB91 test, a word accuracy of 97.2% was obtained using one set of SI models. However, it was noticed that there was often a fairly large variation in word accuracy across the 4 test sets. In particular, the performance was worst on the JUN88 test, which was taken from the speaker dependent data. Therefore, it was decided to look at the errors on the speaker-dependent test data. The recognizer was run on a total of 2700 sentences including the SD-DEV and SD-EVAL data in addition to the SI test data. The phone recognizer was also run on all of the data. (The phone accuracy for this task, without the use of phonotactic constraints was almost 80%.) All of the errors were looked at, comparing the word errors to the recognized phone sequence. As a result the base lexicon was modified, adding alternate pronunciations (such as /goIG/ and /goIn/ for the word "GOING"), allowing certain phones to be optional (for example, the /t/ in "COUNTING"), and correcting errors, and the inter-word phonological rules were extended.

Phonological Rules

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 the phonological rules allows the potential for generalization and extension. However, care must be taken as the alternate pronunciations thus generated can cause errors especially for short words when the rules are applied abusively. The use of phonological rules for the RM task has been previously reported by SRI[1] and AT&T[10]. In the case of AT&T, phonological rules were used only with CI phone models.

Phonological rules were added cautiously, avoiding multiple pronunciations for very short words, deleting phones in short words (2-3 phones), or creating homophones. All the added phones are optional, and phones can be optionally deleted in long words. The phonological rules are applied to the phone graph generated from the baseline lexicon by adding skip arcs to optionally delete phones and adding phone models for alternate pronunciations and inserted phones. The resulting phone model graph which is only 12% larger than the original, is used during training and testing.

Inside words:		*Lexicon*	
Optional phones	COUNTING	kawn{t}IG	
	DIEGO-GARCIA	di{y}ego{#}garsi{y}x	
Alternate pron.	GOING	go{w}I[Gn]	
Between words:	*Rule*		*Example*
"the" alternation	Dx-**V**	→ D[xi]**V**	THE ARTIC
Gemination	t-t	→ {t}t	CLOSEST TO
	@nd-t	→ @n{d}t	OAKLAND TO
Off-glide deletion	aw-m	→ a{w}m	HOW MANY
Stop voicing	k-**V**	→ [kg]**V**	PACIFIC OCEAN
Palatalization	t-y	→ [tC]{y}	LAST YEAR
	d-y	→ [dJ]{y}	DID YORKTOWN
Glide insertion	o-**V**	→ o{w}**V**	TOKYO ARE
	i-**V**	→ i{y}**V**	ME ALL
	R-**V**	→ R{r}**V**	PLUNGER IN

Table 5: Examples of phonological rules. Phones in {} are optional, phones in [] are alternates. **V** stands for vowel and the "-" represents a word boundary.

Some examples of the phonological rules are given in Table 5. These include general rules for well known variants such as palatalization, glide insertion and gemination, as well as rules to handle allophonic variation, using only the reduced phone set. So, instead of having a syllable or word final allophones for the voiceless stops, they are optionally allowed to be replaced with their voiced counterparts. There are more specific rules, such as the deletion of the offglide /w/ in the phone sequence /aw/, as found in the word "how." While this is a fairly general phenomenon, in the context of RM this rule becomes very specific for the word sequences "how much" and "how many."

Figures 1 and 2 illustrate some acoustic differences motivating the use of phonological rules, taken from the training data. The speaker code is given by the three letters in parenthesis. In Figure 1

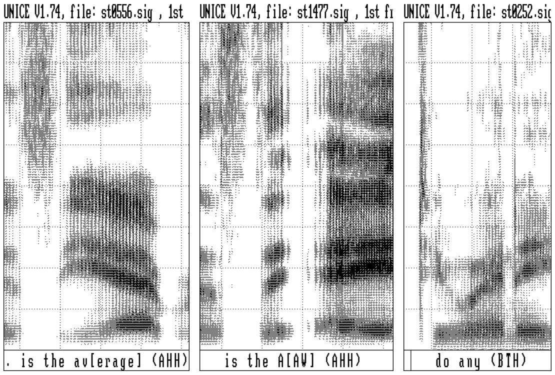

Figure 1: Spectrograms illustrating phonological variation at vowel-vowel boundaries. The scale is 100ms on the horizontal axis and 1kHz on the vertical axis.

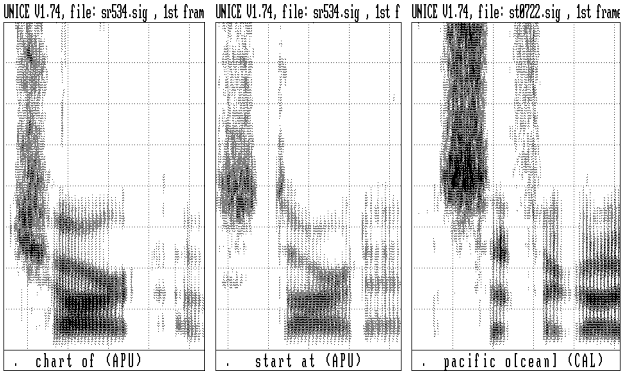

Figure 2: Illustration of stop allophones.

are examples of acoustic realizations at vowel-vowel word boundaries, where it is common to insert either a glide or a glottal stop to mark the boundary. The left most example has a /y/-insertion marking the boundary between in "the average", giving the phone sequence /iy@/. The same speaker, however, uses a glottal stop to mark the boundary in "the AAW", even though the phonetic environment is very similar. The semivowels /r,w/ may be inserted in the same way; an example of a /w/-insertion is seen in the right most spectrogram in the word sequence "do any."

Figure 2 shows some of the variability observed in the realization of stops. The left two spectrograms were taken from the same sentence, and show that even in a similar context, the acoustic realization can be very different. The final /t/ in "chart of" is manifest as a glottal stop, where as the final /t/ in "start at" is flapped. The spectrogram on the right shows that the final /k/ in "pacific ocean" is produced as a /g/. One could argue that this should be considered a speech error, however, the word string is perfectly understood.

Given that even a single speaker may mark phonetic distinctions in different ways, even in a similar phonetic environment, indicates that the use of CD phones as they are typically defined, even if they are word position dependent, will still combine allophones which are acoustically very different. (This distinction was refered to as hard vs soft by Giachin et al.[10].) Therefore, it seems obvious that the use of phonological rules during training will result in purer acoustic models, which should improve the system performance.

The effects of these developmental changes are summarized in Table 6 for four test sets using sex-dependent models. Phase 1 is prior to the use of the speaker-dependent test data, and Phase 2 is after the errors on this data were analysed. It can be seen that the error reduction is between 0% and 20% depending on the test, and that the objective of reducing the difference in performance across tests was acheived.

Test	*JUN88*	*FEB89*	*OCT89*	*FEB91*
Phase 1	4.1	3.2	4.0	2.8
Phase 2	3.3	2.8	3.2	2.9

Table 6: Effect of developmental changes.

Evaluation test: Sep92

The Sep92 evaluation test was run on a Silicon Graphics R3000 workstation, with a liberal prunning threshold. The recognition time per sentence, running the sex-dependent models in parallel, is on the order of 5 min with the word pair grammar and 17 min with no grammar. This is the first time that the system was run without a grammar. The results are summarized in Table 7. (Complete results are reported by NIST, this proceedings.) The word accuracy for the WPG condition was 95.6%. For the NG case, the inter-word phonological rules were not used.

Lexicon	*Corr.*	*Subs.*	*Del.*	*Ins.*	*WErr.*	*SErr.*
WPG	96.0	2.9	1.2	0.4	4.4	25.0
NG[1]	83.2	14.0	2.8	3.2	20.0	70.7

Table 7: Recognition results with a word-pair grammar.

After the evaluation test we attempted to evaluate how much each of the components of the system contributed to the performance for this test set. Unfortunately, it was not possible to undo each component individually as intermediary versions of the system during development were not kept. In addition, most of the changes introduced by the phonological rules affect the sequence of phones in the recognized string. Since the phone contexts modeled are chosen by thresholds based on counts in the training data, any change to the training phone sequence can ultimately affect the particular contexts modeled. Since it seemed more unfair to compare different sets of models, than to evaluate without components used in training, the assessment of the contribution of the various components was made using the same set of models used for the evaluation test which had been trained with all the components. Components were then sequentially removed for the test.

These results are summarized in Table 8. Removing the inter-word phonological rules increases the error rate by about 18%. The removal of the alternate pronunciations had no additional effect on the error rate. Removing the optional phones (which may have been explicitly specified in the lexicon, or added as intra-word phonological rules) increased the error rate by less than 4%. The effect of removing optional within word silences was about the same. Using only one set of SI models gave a word error rate of 5.4%, indicating that for this test the sex-dependent models reduces the error rate by about 20%. Subsequently removing the inter-word phonological rules increased the error by an additional an 11%.

After the Sep92 evaluation, additional performance improvements have been obtained on the development tests using the same

[1] With no inter-word phonological rules.

Condition	*WErr.*
baseline (male/female models, phono. rules)	4.4
- inter-word phonological rules	5.2
- alternate pronunciations	5.2
- optional phones (except silences)	5.4
- optional silences (intra-word)	5.7
SI models (phono. rules)	5.4
- inter-word phonological rules	6.0

Table 8: Assessment of the contribution of some system components on the Sep92 test by sequential removal.

(more liberal) pruning threshold as was used for the official evaluation. The treatment of inter-word silence was also altered, since it was observed that silence could easily be be inserted to take up the slack for poor acoustic matches. These changes resulted in a small error reduction (4%) on the development data: 96.7%(Jun88), 97.5% (Feb89), 96.7% (Oct89), and 97.4% (Feb91).

SUMMARY

In this paper an overview of the speech recognition research at LIMSI has been presented. Our recent work focuses on developing phone-based recognizers that are task-, speaker- and vocabulary-independent so as to be easily adapted to various applications. Phone and word recognition results were reported for French, using data from the BREF corpus. A phone accuracy of 74.2% was obtained using 428 context-dependent phone models, with phonotactic constraints provided by a phone bigram model. The phone recognition results are somewhat superior to those reported for English[16, 21]. This may be simply because French has a smaller number of phonemes, or that the phonemes are less variable due to context. Word recognition for BREF was evaluated on lexicons ranging from 1000 to 10,000 words, for the no-grammar case and with a word-pair grammar. For the no-grammar case the word accuracy was 69.2% with 1000 words and dropped to 49% with 10,000 words. With a word-pair grammar the word accuracy was 87.9% and 86.1% respectively for 1000 and 3000 words.

Phone recognition has also been shown to be powerful for identifying non-linguistic speech features, e.g. sex, language, speaker. Experiments in language identification show that for English sentences, 400 ms sufficed to identify the language as English, whereas for French, 1000 ms were needed to unambiguously identify the language as French. Speaker identification experiments with TIMIT had an identification rate of 99.6%, comparing one utterance from each speaker to models from all 462 training speakers.

The RM Sep92 evaluation system uses a reduced set of 36 phones to represent the lexicon so as to eliminate infrequent phones and to allow more sharing of contexts. CD phone models are used, including cross-word contexts which are position independent. Each phone model is a left-to-right CDHMM with gaussian mixture. Duration is modeled with a gamma distribution per phone model. Phonological rules are used in training to obtain purer acoustic models. The same rules are used in testing so as to allow for unseen events. Separate male and female models were trained and used in parallel. The word accuracy on the Sep92 evaluation test was 95.6%. Average word accuracy on the development tests (1200 sentences: Jun88, Feb89, Oct89, Feb91) is 97.1%.

REFERENCES

[1] M. Cohen, *Phonological Structures for Speech Recognition*, PhD Thesis, U. Ca. Berkeley, 1989.

[2] S.B. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. ASSP*, 28(4), 1980.

[3] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM" NTIS order number PB91-100354.

[4] J.L. Gauvain, "A Syllable-Based Isolated Word Recognition Experiment," *ICASSP-86*.

[5] J.L. Gauvain, "Le système de reconnaisance *AMADEUS*: Principe et algorithmes," LIMSI report, June 1990.

[6] J.L. Gauvain, J.J. Gangolf, "Terminal integrates speech recognition and text-to-speech synthesis", *Speech Technology*, Sept-Oct 1983.

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

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

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

[10] E. Giachin, A.E. Rosenberg, C.H. Lee, "Word Juncture Modeling using Phonological Rules for HMM-based Continuous Speech Recognition,"*Computer Speech & Language*, 5, 1991.

[11] B.H. Juang, "Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains", *AT&T Technical Journal*, 64(6), 1985.

[12] L.F. Lamel, J.L. Gauvain, "Experiments on Speaker-Independent Phone Recognition Using BREF," *ICASSP-92*.

[13] L.F. Lamel, J.L. Gauvain, "Multi-lingual Speech Recognition at LIMSI," Presented at the 1st Intl. Workshop on Speech Translation, Warden, Germany, Oct. 18-20, 1992.

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

[15] C.H. Lee, L.R. Rabiner, R. Pieraccini and J.G. Wilpon, "Acoustic modeling for large vocabulary speech recognition," *Computer Speech & Language*, 4, 1990.

[16] K.F. Lee, H.W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," *IEEE Trans. ASSP*, 37(11), 1989.

[17] K. Matrouf, J.L. Gauvain, F. Néel, J. Mariani, "Adapting Probability-Transitions in DP Matching Process for an Oral Task-Oriented Dialogue," *ICASSP-90*.

[18] D. Paul, J. Baker, "The Design for the Wall Street Journal-based CSR Corpus" *Proc. DARPA Speech & Nat. Lang. Workshop*, Feb. 1992.

[19] G.M. Quénot, J.L. Gauvain, J.J. Gangolf, J. Mariani, "A Dynamic Programming Processor for Speech Recognition", *IEEE J. of Solid-State Circuits*, 24(2), 1989.

[20] L.R. Rabiner, B.H. Juang, S.E. Levinson, M.M. Sondhi, "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities," *AT&T Technical Journal*, 64(6), 1985.

[21] T. Robinson, F. Fallside, "A recurrent error propogation network speech recognition system," *Computer Speech & Language*, 5, 1991.