ELSEVIER

Speech Communication 15 (1994) 21–37

SPEECH COMMUNICATION

Speaker-independent continuous speech dictation ☆

J.L. Gauvain *, L.F. Lamel, G. Adda, M. Adda-Decker

LIMSI-CNRS, BP 133, 91403 Orsay cedex, France

Received 16 February 1994; revised 5 August 1994

Abstract

In this paper we report on progress made at LIMSI in speaker-independent large vocabulary speech dictation using newspaper-based speech corpora in English and French. The recognizer makes use of continuous density HMMs with Gaussian mixtures for acoustic modeling and n-gram statistics estimated on newspaper texts for language modeling. Acoustic modeling uses cepstrum-based features, context-dependent phone models (intra and interword), phone duration models, and sex-dependent models. For English the ARPA *Wall Street Journal*-based CSR corpus is used and for French the BREF corpus containing recordings of texts from the French newspaper *Le Monde* is used. Experiments were carried out with both these corpora at the phone level and at the word level with vocabularies containing up to 20,000 words. Word recognition experiments are also described for the ARPA RM task which has been widely used to evaluate and compare systems.

Zusammenfassung

In diesem Beitrag beschreiben wir Fortschritte in der Entwicklung eines sprecherunabhängigen Spracherkennungssystems für großen Wortschatz, welches mit (gesprochenem und geschriebenem) Datenmaterial von Zeitungartikeln trainiert wurde. Die akustische Modellierung des Spracherkennungssystems besteht aus Mischungen kontinuierlicher gauß'schen Dichten in Hidden Markov Modellen (HMM). Die Modellierung der (geschriebenen) Sprache beruht auf statistischen n-grams, deren Wahrscheinlichkeiten aus einer Datenbasis bestehend aus Zeitungsartikeln geschätzt wurden. Was die akustischen Modelle betrifft, verwenden wir Cepstrum Parameter in kontextabhängigen Phonmodellen, mit zeitlicher Modellierung und geschlechtspezifischen Modellen. Für die englische Sprache benutzen wir die ARPA *Wall Street Journal* Datenbasis und für die französische, die BREF Daten, die gesprochene Texte der französischen Zeitung *Le Monde* enthalten. Für beide Sprachen wurden Experimente auf Phonem- und Wortbasis durchgeführt. Der Wortschatz besteht aus bis zu 20K Wörter. Weiterhin präsentieren wir Ergebnisse auf der ARPA RM Datenbasis, da diese weltweit zur Bewertung von Spracherkennungssystemen verwendet wurde.

☆ This paper is based on a communication presented at the ESCA Conference Eurospeech-93 and has been recommended by the Eurospeech-93 scientific program committee.

* Corresponding author.

SSDI 0167-6393(94)00024-7

Résumé

Nous présentons dans cet article les avancées réalisées au LIMSI sur la reconnaissance de parole continue de grand vocabulaire, indépendante du locuteur dans une application de dictée de textes. Le système utilise des modèles de Markov cachés à densités continues au niveau acoustique, et des modèles de langage n-grammes au niveau syntaxique. La modélisation acoustique repose sur une analyse cepstrale du signal vocal, des modèles de phones en contexte (inter- et intramot) dépendant du genre du locuteur, et des modèles de durée phonémique. Nous avons utilisé, pour la langue anglaise, le corpus de parole continue ARPA-WSJ contenant des enregistrements de textes lus extraits du *Wall Street Journal*, et, pour la langue française, le corpus BREF contenant des enregistrements de textes lus extraits du journal *Le Monde*. Les performances du système de reconnaissance, mesurées au niveau phonétique et au niveau mot sont données sur ces deux corpus pour des vocabulaires contenant jusqu'à 20.000 mots. Nous donnons également pour référence les résultats obtenus sur le corpus ARPA-RM qui a été très largement utilisé pour évaluer et comparer des systèmes de reconnaissance de parole.

Keywords: Continuous speech recognition; Word recognition; Phone recognition; Speaker-independent; Large vocabulary; Dictation

1. Introduction

An outstanding challenge in continuous speech recognition research is to develop recognizers that are task-, speaker- and vocabulary-independent so as to be easily adapted to various applications. In this paper we report on recent efforts at LIMSI in large vocabulary, speaker-independent continuous speech recognition in English and French, and address some language-dependent issues. Three corpora have been used to carry out the experiments: the ARPA Resource Management corpus (RM) (Price et al., 1988), the ARPA *Wall Street Journal*-based CSR corpus (WSJ) (Paul and Baker, 1992), and the BREF *Le Monde*-based corpus (Lamel et al., 1991). All three corpora contain large amounts of read speech material from a large number of speakers, recorded under similar conditions (8 kHz bandwidth, close-talking microphone, read-speech). WSJ and BREF also have associated text materials which are used as a source for statistical language modeling. For these two corpora, experiments have been carried out at the phone level and at the word level with comparable size lexicons and test perplexities. The recognizer was evaluated in the September 1992 ARPA continuous speech recognition evaluation on the 1000-word Resource Management task (Pallett and Fiscus, 1992) and also in the ARPA Wall Street Journal evaluation in November 1992 (Pallett et al., 1993). [1]

This paper is organized as follows. In the next section the recognizer is described, emphasizing the characteristics which are different from other HMM-based systems. The following three sections present experiments on each of the RM, WSJ, and BREF tasks, including descriptions of the corpus and task specific details. For WSJ and BREF, phone recognition and word recognition results are presented. The final section points out links observed between phone and word recognition, discusses some of the problems encountered in the dictation task, and highlights some language-dependent differences at both the phone and the word level.

2. Recognizer overview

The recognizer makes use of n-gram statistics for language modeling and of continuous density

[1] The recognizer has since also been evaluated in the ARPA November 1993 Wall Street Journal evaluation (Pallett et al., 1994; Gauvain et al., 1994b). For coherency the experimental results given in this paper for the Nov93 system are on the Nov92 test data. We would like to point out, however, that word error on the Nov92 test data is lower than that obtained on the Nov93 test and on several sets of development test data.

HMMs with Gaussian mixtures for acoustic modeling. A time-synchronous graph-search strategy (Ney, 1984) which includes intra- and interword context-dependent phone models, intra-and interword phonological rules, phone duration models, gender-dependent phone models is used with a bigram-backoff language model (Katz, 1987). When a trigram LM is used, a second acoustic decoding pass is carried out making use of a word graph generated with the bigram LM (Gauvain et al., 1994b). 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 then replaced by a phone graph obtained by concatenation of the phone models of the word according to its phone transcription in the lexicon.

2.1. Acoustic front end

A feature vector is computed every 10 ms on an 8 kHz bandwidth. For each frame (30 ms window), a 15 channel Bark power spectrum is obtained by applying triangular windows to the DFT output. The cepstrum coefficients are then computed using a cosine transform (Davis and Mermelstein, 1980). The feature set includes 15 Bark-frequency scale cepstrum coefficents with their first and second order derivatives (Δ and $\Delta\Delta$ cepstrum) as well as the log-energy and its first and second order derivatives. Some comparative experiments were carried out using 4 kHz and 8 kHz bandwidths (see Tables 5 and 8), as well as with different feature vectors (see Tables 5, 6, and 8).

2.2. Acoustic models

The acoustic models are sets of context-dependent (CD) phone models, which include both intraword and cross-word contexts, but are position independent. The contexts to be modeled are automatically selected based on their frequencies in the training data. The CD units include triphone models, right-context phone models, left-context phone models, and context-independent phone models. Each phone model is a left-to-right continuous density HMM with Gaussian mixture observation densities (typically 32 components). The covariance matrices of all the Gaussian components are diagonal. Since phone duration is not adequately modeled with a three state Markov chain, a separate duration density is associated with each phone model. Duration is thus modeled with a gamma distribution per phone. As proposed by Rabiner et al. (1985), the HMM and duration parameters are estimated separately and combined in the recognition process during the Viterbi search. Maximum a posteriori (MAP) estimators are used for the HMM parameters (Gauvain and Lee, 1992) and moment estimators for the gamma distributions. The use of a priori knowledge in the acoustic parameter estimation process has been shown to be particularly effective (Lee et al., 1990; Gauvain and Lee, 1994) and can be used to easily adapt SI models to gender-dependent (or speaker-dependent) models. Separate male and female models are thus obtained to more accurately model the speech data. Context-dependent phone modeling is able to account for a large part of coarticulation. Nonetheless, in part because the CD models are position independent, for a given triphone there can be rather different acoustic realizations. This problem has been addressed by using as many mixture components as possible and by introducing phonological rules (see below).

2.3. Lexicon

The lexicon is represented phonemically, with different lexicons for each task. The phone sets for RM and WSJ are given in Table 1 and the phone set of BREF is given in Table 2. The RM lexicon has 990 lexical entries. For WSJ and BREF, test lexicons containing 5,000 and 20,000 words are used. The lexicons include alternate pronunciations for some of the words, and allow some of the phones to be optional. [2] For each word the baseform transcription is used to generate a pronunciation graph to which word-internal

[2] About 10% of the lexical entries have multiple transcriptions. For BREF, this count does not include alternate transcriptions due to final optional phonemes marking possible liaisons. Including these raises the number of entries with multiple transcriptions to almost 40%.

Table 1
Phone symbol sets for English. For RM a reduced set of 36 phones are used. For WSJ 46 phones are used

Phone		Example word	Phone		Example word
RM	WSJ		RM	WSJ	
Vowels			*Fricatives*		
i	i	beet	s	s	sue
I	I	bit	z	z	zoo
e	e	bait	S	S	shoe
E	E	bet		Z	measure
@	@	bat	f	f	fan
ʌ	ʌ	but	v	v	van
a	a	bott	T	T	thin
c	c	bought	D	D	that
o	o	boat	*Affricates*		
u	u	boot		C	cheap
	U	book	J	J	jeep
R	R	bird	*Plosives*		
Diphthongs			p	p	pet
	Y	bite	t	t	tat
	O	boy	k	k	cat
	W	bout	b	b	bet
Reduced vowels			d	d	debt
x	x	about	g	g	get
	|	dated	F		butter
	X	butter	*Nasals*		
Semivowels			m	m	met
l	l	led	n	n	net
r	r	red	G	G	thing
w	w	wed	*Syllabics*		
y	y	yet		L	bottle
h	h	hat		M	bottom
·	·	silence		N	button

Table 2
The 35-phone symbol set for French

Phone	Example	Phone	Example
Vowels		*Fricatives*	
i	lit	f	fou
e	blé	v	vin
E	sel	s	sot
y	suc	z	zèbre
X	leur	S	chat
x	petit	Z	jour
@	feu	*Plosives*	
a	patte, pâte	p	pont
c	sol	b	bon
o	saule	t	ton
u	fou	d	don
Nasal vowels		k	cou
I	brin, brun	g	gond
A	chant	*Nasals*	
O	bon	m	motte
Semivowels		n	note
h	lui	N	digne
w	oui	·	silence
j	yole		
l	la		
r	rond		

phonological rules are optionally applied during training and recognition to account for some of the phonological variations observed in fluent speech. Examples of some typical word internal phonological rules are given in Fig. 1 using the phone symbol set given in Table 1. These include the optional /t/ in COUNTING and the phonological variant of word-final "*ing*" (/G/) realized as "*in*" (/n/). The examples of cross-word phonological rules are discussed in Section 2.6.

2.4. Language model

For RM, the standard deterministic word-pair grammar was used. For WSJ and BREF, bigram and trigram-backoff (Katz, 1987) language models (LMs) were estimated on the training text material. In the WSJ baseline LMs the backoff mechanism is used for unobserved word sequences and bigrams observed only once (Paul and Baker, 1993). To give an idea of the LM size, the WSJ 20k open vocabulary nvp bigram LM has 1.5M bigrams and the 20k open vocabulary nvp trigram

Within-word phonological rules:

		Lexicon
Optional phones	COUNTING	kawn{t}|G
	ACCORDINGLY	xkcrd|G{g}li
Alternate pron.	GOING	go{w}|[Gn]

Cross-word phonological rules:

	Rule	*Example*
"the" alternation	Dx-V → D[xi]V	THE APPLE
Gemination	t-t → {t}t	CLOSEST 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 YOU
Glide insertion	o-V → o{w}V	SO ARE
	i-V → i{y}V	ME ALL

Fig. 1. Examples of lexical representation and phonological rules. Some of the cross-word phonological rules are specific to the RM task. Phones in {} are optional, phones in [] are alternates. **V** stands for vowel and the "-" represents a word boundary. The phone symbol set is given in Table 1.

LM has 3.1M trigrams. For BREF, the 20k bigram LM has 1.2M bigrams and the 20k trigram has 3.2M trigrams. The LM size can be substantially reduced by relying more on the backoff, i.e., by slightly increasing the minimum number of required word sequence observations needed to include the n-gram.

For the phone recognition experiments, phone 2-gram probabilities are used to provide phonotactic constraints. Since only orthographic transcriptions are provided for the WSJ corpus and all but a small portion of the BREF corpus (Gauvain and Lamel, 1992), phone transcriptions were automatically generated in the following fashion. A Markov chain corresponding to all the possible phone strings for the given sentence is generated based on the phone transcriptions in the associated lexicon and the phonological rules. Forced alignment is then performed with the speech signal and the best aligned string is considered to be the reference phone transcription, [3] which is then used to train the phone LMs. These phone 2-gram probabilities are used to provide phonotactic constraints corresponding to the between phone model transition probabilities.

2.5. Decoding

The recognizer uses a time-synchronous graph-search strategy (Ney, 1984) with a bigram-backoff language model (Katz, 1987) which can be used to generate, in addition to the best solution, a word graph. When a trigram LM is used, a second acoustic decoding pass is performed making use of this word graph (Gauvain et al., 1994b). A classical beam search strategy is used to limit the search space. Both passes incorporate intra- and interword CD phone models, intra- and interword phonological rules and phone duration models. In the first pass, the backoff component of the bigram-backoff language model is efficiently implemented with a tree organization of the lexicon which significantly reduces the number of connections between words in the search graph. This is important because the number of interword connections can be quite large due to the use of interword triphones. It should be noted that this decoding strategy based on two forward passes can in fact be implemented in a single forward pass. A two pass solution has been chosen because it is conceptually simpler, and also due to memory constraints. In terms of computation, the second pass is carried out in only a fraction (about 1/5) of the time of the first pass.

During phone recognition the male and female models are run in parallel, and the output with the highest likelihood is chosen. For the word recognition tests, gender-selection is first performed for each sentence using phone-based ergodic HMMs (Lamel and Gauvain, 1993c). The word recognizer is then run using the set of models corresponding to the identified sex.

2.6. Phonological rules

Phonological rules are used to allow for some of the phonological variations observed in fluent speech. 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 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. A pronunciation graph is generated for each word from the baseform transcription to which word internal phonological rules are applied. In forming the word network, word boundary phonological rules are applied at the phone level to take into account interword phonological variations, such as palatalization, voicing assimilation, or glide insertion for English. Some examples of cross-word phonological rules are given in Fig. 1. The same mechanism has been used to handle liaisons, mute-e, and final consonant clus-

[3] While this method of defining the reference phone transcription may be considered a bit optimistic, our experience has shown that the difference in phone error using automatically generated and manually corrected transcriptions is very small. On a set of 200 sentences, the overall phone error increased by an absolute value of 0.3% (Lamel and Gauvain, 1993b).

ter reduction for French. Fig. 2 gives some examples taken from the RM training data illustrating acoustic differences occurring at vowel–vowel word boundaries which can be efficiently dealt with using phonological rules. The RM speaker code is given by the three letters in parentheses. It is common to mark vowel–vowel boundaries by inserting a glide or making a glottal stop. 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*" (middle example), even though the phonetic environment is very similar. In the example on the right, a /w/ is inserted.

2.7. System development

Much of system development has been carried out by performing phone recognition instead of word recognition, in order to reduce the computational requirements and speed up the development process. We have shown that improvements in phone accuracy are directly indicative of improvements in word accuracy when the same phone models are used for recognition (Lamel and Gauvain, 1993b). This has allowed us to evaluate many alternatives for the front-end and the acoustic models. Phone recognition provides the added benefit that the recognized phone string can be used to understand errors in word recognition and problems with the lexical representation.

In this paper we report results in phone recognition and word recognition using various sets of CD phone models. Since the CD units to be modeled are selected based on their frequency in the training corpus, the size of the model set can be controlled by varying the minimum number of occurrences necessary to model a given context, so as to match the number of parameters of the recognizer to the available training data. As will be demonstrated in the experimental sections (Sections 4 and 5), recognition performance can be improved by increasing the number of contexts modeled, provided that there are sufficient occur-

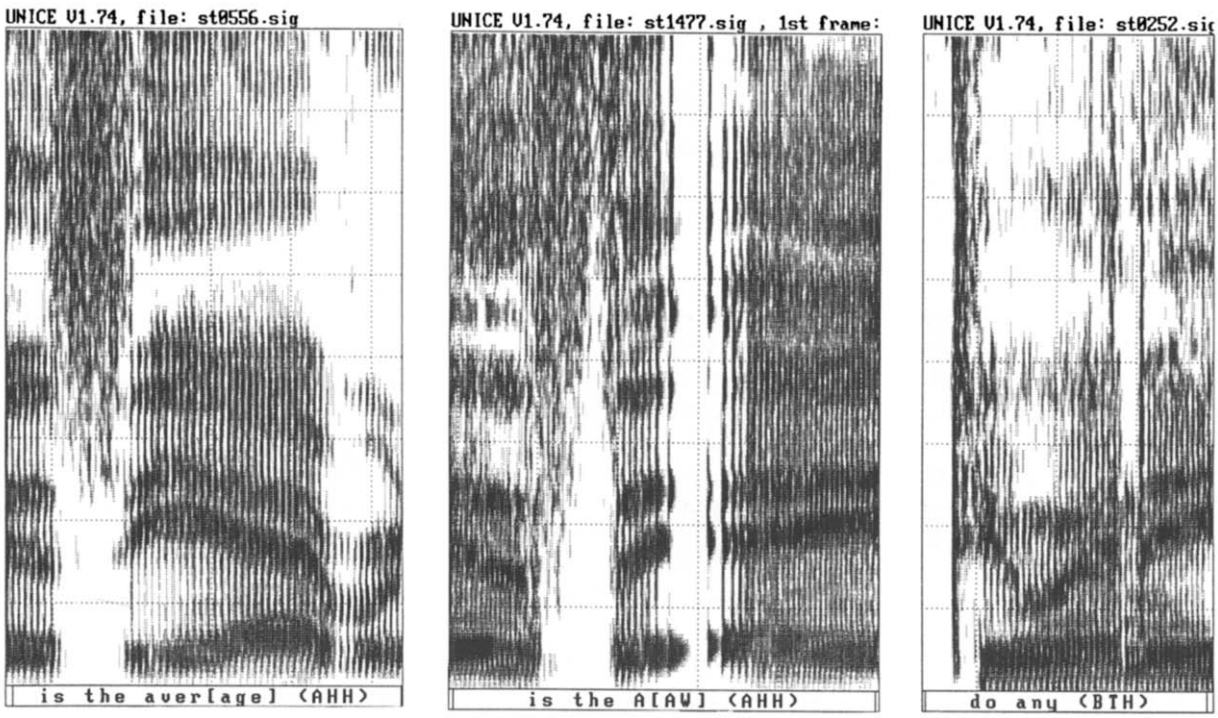

Fig. 2. Spectrograms from the RM training data illustrating phonological variation at vowel–vowel boundaries. The scale is 100 ms on the horizontal axis and 1 kHz on the vertical axis.

rences of these contexts in the training data. In practice we have found that a minimum number of 250 occurrences are needed to accurately model a given context and that reducing this value, thereby modeling more contexts, typically does not give any further improvement in phone or word accuracy. The RM task is an exception to this rule, where due to the use of a weak language model (standard word-pair grammar) it can be advantageous to increase the number of CD units, particulary the cross-word units. For this task an optimum value for the minimum number of required occurrences was observed to be around 25.

3. Experiments using RM

The ARPA Resource Management speech corpus (Price et al., 1988) is a corpus of read speech with a medium size vocabulary (1000 words) designed to provide speech data for evaluation of continuous speech recognizers and has been widely used in comparative evaluations. We include these results here in order to allow the comparison of our system to other recognizers developed worldwide. In these experiments the standard set of 3990 sentences (SI-109) has been used to train two sets of 2274 CD phone models, from the male and female speakers' data. The standard word-pair grammar (perplexity 60) was used.

The JUN88, FEB89, and OCT89 SI test sets were used as development data to evaluate various alternatives for the front end, the lexical representation, and the phonological rules, and to estimate some parameter values such as the word insertion penalty. These data sets were then complemented with SD-DEV and SD-EVAL data (for a total of 2700 sentences) and the most common errors were analyzed and used to add alternate pronunciations to the baseline lexicon and to create some task-specific phonological rules. This error analysis was not only based on the word recognizer output but also on the phone recognizer output. The FEB91 test data was reserved for evaluation at the end of each development cycle.

The RM lexicon is represented with a set of 36 phones, as given in Table 1. This reduced set was used primarily to eliminate infrequent phones for which there was insufficient training data, and to provide a means of better sharing contexts. In doing so, more data is available to train the remaining models, and the number of potential triphone contexts is reduced. The infrequent phones /Z,U/ were eliminated and replaced by another "close" phone. The diphthongs /Y,O,W/ were represented by a sequence of phones. Allophonic distinctions such as the syllabics /L,M,N/, the context-dependent distinction between the two schwas /x, |/, and the stress difference between /X,R/, were not 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.

The lexicon provides alternate pronunciations for about 10% of the words. For example, the word MONTICELLO has the pronunciations /mantxsElo/ and /mantxtSElo/, and the /t/ in COUNTING (/kawn{t}IG/) is optional. Intra- and interword phonological rules are optionally applied during training and recognition. The use of phonological rules for the RM task has been previously reported by SRI (Cohen, 1989) and AT&T (Giachin et al., 1991). In the case of AT&T, phonological rules were used only with CI phone models. A single speaker may mark phonetic distinctions in different ways even in similar phonetic environments. This means that the use of CD phones as they are typically defined, combines allophones which can be acoustically very different. The use of phonological rules during training should result in purer acoustic models, and thus improve the system performance.

Some examples of phonological rules are given in Table 1. These include general rules for well known variants such as palatalization, glide insertion and gemination, as well as rules to handle allophonic variation. For the RM task some additional phonological rules are used. For example, since the CD models are position independent, there are no syllable-final or word-final allophones for the voiceless stops. These stops are

Table 3
Word recognition results on the ARPA-RM-SI corpus with a WP grammar of perplexity 60

ARPA test	Corr.	Subs.	Del.	Ins.	WErr.
JUN88	97.1	2.5	0.4	0.4	3.3
FEB89	97.7	1.7	0.5	0.2	2.5
OCT89	97.0	2.2	0.9	0.3	3.3
FEB91	97.7	1.9	0.4	0.3	2.6
SEP92 [a]	96.0	2.9	1.2	0.4	4.4

[a] Official ARPA SEP92 evaluation results.

Table 4
Assessment of the contribution of some system components on the Sep92 test by sequential removal

Condition	WErr.
Baseline (male/female models, phono. rules)	4.4
– interword phonological rules	5.2
– alternate pronunciations	5.2
– optional phones (except silences)	5.4
– optional silences (intraword)	5.7
SI models (phono. rules)	5.4
– interword phonological rules	6.0

therefore optionally allowed to be replaced by their voiced counterpart. A more specific rule allows for the deletion of the offglide /w/ in the phone sequence /aw/, in certain contexts. 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".

The developmental changes based on the error analysis provided an 18% reduction on the word error rate measured on the development data (Lamel and Gauvain, 1992). Results on the last 5 ARPA tests are reported in Table 3. After the Sep92 ARPA test, the contribution of the system components to the performance on the Sep92 test data was assessed as shown in Table 4 by successively removing components of the system. These results indicate that the interword phonological rules and the sex-dependent models had the largest influence in reducing the word error.

4. Experiments using WSJ

The ARPA WSJ corpus (Paul and Baker, 1992) was designed to provide general-purpose speech data with large vocabularies. Text materials were selected to provide training and test data for 5K and 20K word, closed and open vocabularies, with both verbalized (vp) and non-verbalized (nvp) punctuation. The recorded speech material supports both speaker-dependent and speaker-independent training and evaluation. In these experiments, data from the WSJ0 and WSJ0/WSJ1 corpora were used. The standard WSJ0 SI-84 training data include 7240 sentences from 84 speakers (42m, 42f). The standard WSJ0/WSJ1 SI-284 training data contains 37,518 sentences from 284 speakers. For word recognition, the standard WSJ language models trained on the 37M word normalized training text material were used.

4.1. Phone recognition

The phone accuracy was assessed on the non-verbalized and verbalized punctuation Feb92 pilot evaluation material containing 200 sentences from 10 speakers (6m/4f) for each condition. Since there are no associated phone transcriptions for this data, a phone transcription was determined by performing segmentation as described in Section 2.4. A set of 46 phones were used, consisting of 21 vowels, 24 consonants, and silence as given in Table 1. Phonotactic constraints were provided by a phone bigram estimated on automatically generated phone labels of the training data. The phone perplexity is 17.5 for the nvp test data and 15.1 for the vp test data.

Table 5
Phone recognition results for WSJ Feb92-5k pilot evaluation material using 46 phones and phone bigram. All model sets were trained with WSJ0 training material, except 3306m, where the WSJ0/WSJ1 training material was used

Conditions	Corr.	Subs.	Del.	Ins.	Err.
4 kHz, Δ, 493m, nvp	71.3	21.4	7.3	5.2	**33.9**
8 kHz, Δ, 493m, nvp	74.8	18.7	6.5	4.9	**30.1**
8 kHz, ΔΔ, 493m, nvp	77.0	17.1	5.9	4.6	**27.6**
8 kHz, ΔΔ, 884m, nvp	78.9	16.2	4.9	4.8	**25.9**
8 kHz, ΔΔ, 1619m, nvp	79.3	16.2	4.5	5.0	**25.7**
8 kHz, ΔΔ, 3306m, nvp	85.5	11.6	2.9	4.7	**19.2**
8 kHz, ΔΔ, 3306m, vp	87.8	11.6	2.2	4.1	**16.5**

nvp: non-verbalized punctuation, vp: verbalized punctuation.

Experimental results for the Feb92 pilot test data are given in Table 5 where silences have been removed prior to scoring. For each size of CD model, separate male and female models were trained, and used in parallel during recognition. The recognized string is that associated with the model set having the highest likelihood. The first two entries compare the phone accuracy for 4 kHz and 8 kHz bandwidths with a feature vector containing the cepstrum and the Δ cepstrum, using 493 CD models trained with the WSJ0 training material. An absolute reduction in the phone error of almost 4% is obtained with the larger bandwidth [4]. Increasing the size of the feature vector to include the ΔΔ cepstrum gave an additional absolute error reduction of 2.5% with the same model set. The next entry shows that when the number of CD models is increased to 884, the absolute error is reduced by 1.7%. For the 884 model set a minimum number of 250 occurrences was required to model a context. Reducing this threshold, and thereby increasing the number of models to 1619 gives only a small error reduction of 0.2%. In numerous cases we have observed that reducing the minimum number of occurrences below 250 does not give significant improvement in recognition accuracy. Therefore, in order to train a larger set of acoustic models, it is necessary to have additional speech data.

The last two entries give results using the combined WSJ0/WSJ1 training data. Using a set of 3306 models, the phone error is seen to be reduced by about 25% over the 1619 models trained with the WSJ0 training data. The phone accuracy on the Feb92 vp test data using the same set of 3306 CD models is 16.5%. This higher accuracy can be attributed to the frequent occurrence of the phones in the punctuation words (particularly *period* and *comma*), which are both well modeled and well recognized.

[4] Similar improvements in phone accuracy with a larger bandwidth were observed also for the TIMIT corpus (Lamel and Gauvain, 1993a,b), indicating that the frequency range carries relevant information for American English.

4.2. Word recognition

Two series of word recognition experiments investigating issues in acoustic modeling and language modeling are reported in this section. In the first set of experiments, the acoustic models were trained on the WSJ0 training data and a bigram-backoff language model was used. In the second set of experiments, the combined WSJ0/WSJ1 training data was used with both bigram and trigram language models. The standard bigram and trigram-backoff language models provided by Lincoln Labs (Paul and Baker, 1992) were estimated on the 37 million word standardized WSJ text material. The lexicon is represented using the set of 46 phones given in Table 1. The pronunciations were obtained from various existing lexicons (TIMIT, Pocket and Moby), and 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. In manually verifying the pronunciations, optional and/or alternate phonemes were added.

Phonological rules were optionally applied during training and test. For the present, only well known phonological rules have been incorporated in the system. These rules include both word-internal rules and interword rules as previously shown in Fig. 1.

The first set of experiments made use of acoustic models trained on the WSJ0 training

Table 6
Word recognition results on the Nov92 test data, with acoustic models trained on the WSJ0 corpus and a probabilistic grammar (bigram) estimated on WSJ text data

Conditions [a]	Corr.	Subs.	Del.	Ins.	Err.
493m, Δ, 5k, nvp[b]	91.8	6.9	1.3	1.5	**9.7**
493m, Δ, 5k, vp[b]	93.6	5.5	0.9	1.4	**7.8**
884m, ΔΔ, 5k, nvp	94.1	5.2	0.7	1.0	**6.9**
884m, ΔΔ, 5k, vp	94.5	4.7	0.7	1.1	**6.5**
884m, ΔΔ, 20k, nvp	88.3	10.1	1.5	2.0	**13.6**
884m, ΔΔ, 20k+, nvp	86.8	11.7	1.5	2.7	**15.9**

[a] 5k: 5000 word lexicon, 20k: 20,000 word lexicon, 20k+ 20,000 word lexicon with open test, nvp: non verbalized punctuation, vp: verbalized punctuation.
[b] Official ARPA NOV92 evaluation results.

data and bigram LMs. This system was evaluated in the Nov92 ARPA evaluation test for the 5k closed vocabulary (330 sentences from 8 speakers) using the standard bigram language models. The official reported results are given in the first two lines of Table 6 using the same sets of sex-dependent 493 CD models for the nvp and vp conditions, without the second derivative of the cepstral coefficients. Increasing the number of CD models and the number of features, reduced the relative error rate by about 20% over the system used for the Nov92 evaluation. Results of this latter system on the Nov92 nvp 64k test data (333 sentences from the same 8 speakers) are also given in Table 6 for 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 (Paul and Baker, 1992).) 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 (nvp tests). The higher error rate with the 20k open (20k +) lexicon can be largely attributed to the out-of-vocabulary words, which account for almost 2% of the words in the test sentences. On average there are 1.2 errors made for each OOV word, implying that in most cases the OOV word is simply replaced by another word, and sometimes it is replaced by a sequence of two or more words.

One problem in using a bigram LM for large vocabularies is that the number of interword connections in the search graph is very large. We investigated the effects of reducing the size of the bigram LM by relying more on the backoff, taking advantage of our lexicon tree organization of the backoff component. Using a count threshold of 4 occurrences reduces the bigram size by 53% and gives a word error of 7.2% on the 5k-nvp test. This is only a slight increase in the error compared to the 6.9% obtained with a threshold of 1 (standard bigram).

In the second series of experiments, the effects of using substantially more training data were investigated. The results obtained using the combined WSJ0/WSJ1 training material are given in Table 7 with bigram and trigram language models. Using the additional acoustic training data reduced the word error by about 30% for both vocabulary sizes. When a trigram LM was used in the second pass (see Section 2.5), the word error was reduced by another 35% on the 5k test data and by 17% on the 20k + test data. The trigram is necessarily less effective for the 20k + data due to the presence of OOV words that are not modeled. These OOV words occured in 26% of the sentences.

Table 7
Word recognition results on the Nov92 nvp test data, with 3306 acoustic models (ΔΔ) trained on the combined WSJ0/WSJ1 corpus and probabilistic grammars (bigram or trigram) estimated on WSJ text data

Conditions [a]	Corr.	Subs.	Del.	Ins.	Err.
5k, bg	96.0	3.6	0.3	0.9	**4.8**
5k, tg	97.7	2.1	0.2	0.8	**3.1**
20k+, bg	91.6	7.6	0.8	2.6	**11.0**
20k+, tg	93.2	6.2	0.6	2.3	**9.1**

[a] 5k: 5000 word lexicon, 20k+: 20,000 word lexicon with open test.

5. Experiments using BREF

The French BREF corpus contains more than 100 hours of read-speech material, from 120 speakers (55m/65f) (Lamel et al., 1991). The text materials were selected verbatim [5] from the French newspaper *Le Monde*, so as to provide a large vocabulary (over 20,000 words) and a wide range of phonetic environments (Gauvain et al., 1990). Containing 1115 distinct diphones and over 17,500 triphones, BREF can be used to train vocabulary independent phone models. The text material was read without verbalized punctuation. Two sets of acoustic training data were used in these experiments: the si-3k training material containing 2770 sentences from 57 speakers

[5] This is in contrast to the WSJ0 corpus, where the prompts were normalized prior to presentation to the speaker so as to fix the pronunciation of items such as numbers and dates that are subject to variation. This constraint was relaxed when the WSJ1 corpus was recorded.

(28m/29f) and the si-38k training material data containing 38,550 utterances from 80 speakers. While we have previously reported word recognition results using 4M words of *Le Monde* text as language model training material (Gauvain et al., 1994a), in these experiments a larger corpus of 38M words is used. The use of comparable training materials (about 38k utterances and 37M words of text) facilitates the comparison of speech recognition performance in English and French on a similar task.

5.1. Phone recognition

We have previously reported a phone error of 21.3% using the si-3k training data and a test set of 93 sentences from 8 speakers (4m/4f) (Lamel and Gauvain, 1993b). In this paper we report results on the 5k portion of the Feb94-dev test data, containing 25 sentences from each of 8 speakers (5m/3f). The prompts of the test material are distinct from the training texts and have a phone perplexity of 16.1. Phone transcriptions of these utterances were automatically generated using the set of 35 phones including 14 vowels, 20 consonants, and silence, given in Table 2. The phone bigram was estimated on automatically generated phoneme transcriptions of a portion of the training text material in the *Le Monde* corpus (Prouts, 1980).

Phone recognition results on the Feb94-dev 5k data are given in Table 8 using gender-specific sets of CD models. Silences were removed prior to scoring. The first three entries provide phone recognition results for different model sizes using the si-3k training material. By comparing the first two entries it can be seen that increasing the bandwidth to 8 kHz from 4 kHz did not improve the phone error. Thus, in contrast to the observation for WSJ, increasing the bandwidth for French is not particularly useful. However, since slight reductions in the phone error (about 0.2%) have been consistently observed on other test sets, the larger bandwidth was used in the remaining experiments. Including the $\Delta\Delta$ cepstrum in the feature vector reduced the phone error to 20.5%. By increasing the training data to 38k sentences, more CD contexts were able to be modeled, and the phone error was reduced. With 1747 models a phone error of 14.4% was obtained, and with 2964 models the phone error is 13.5%. This high performance in phone recognition has also been obtained on other test data from the BREF corpus, including other test data from the same and from different speakers.

Table 8
Phone recognition results for BREF on Feb94-dev 5k test data using 35 phones, si-3k or si-38k training data, a phone bigram

Conditions	Corr.	Subs.	Del.	Ins.	Err.
428m, 4 kHz, Δ, si-3k	81.6	13.6	4.8	4.3	**22.7**
428m, 8 kHz, Δ, si-3k	81.4	13.9	4.6	4.1	**22.7**
428m, 8 kHz, $\Delta\Delta$, si-3k	83.5	12.6	3.9	4.0	**20.5**
1747m, 8 kHz, $\Delta\Delta$, si-38k	88.9	9.1	2.0	3.3	**14.4**
2964m, 8 kHz, $\Delta\Delta$, si-38k	89.7	8.5	1.9	3.2	**13.5**

5.2. Word recognition

In order to compare word recognition of French with that of English, similar vocabularies, language models and test sets were selected. As for the WSJ task, two vocabularies were used for the recognition experiments, corresponding to the 5k and 20k most common words in the *Le Monde* texts. The base lexicon, represented with the same 35 phonemes as used in the phone recognition experiments, was obtained using text-to-phoneme rules (Prouts, 1980), and was extended to annotate potential liaisons and pronunciation variants. The phone models used here correspond to the best configuration in the phone recognition experiments for each training data set (si-3k or si-38k).

For each vocabulary, language models were estimated on the normalized training text materials from *Le Monde*. Normalization of the text material entailed a processing rather different from the pre-treatment of the WSJ texts (Paul and Baker, 1992). The main differences are in the treatment of upper and lower case, compound words and abbreviations. In WSJ case is not distinctive, whereas in BREF the distinction between the cases is kept when the upper case designates a distinctive graphemic feature, but

not when the upper case is simply due to the fact that the word occurs at the beginning of the sentence. Thus, the first word of each sentence has been semi-automatically verified to determine if a transformation to lower case was needed. Special treatment is also needed for the symbols hyphen (-), quote ('), and period (.) which can lead to ambiguous separations. For example, the hyphen in compound words like "beaux-arts" and "au-dessus" is treated as word-internal. It may also be associated with the first word as in "ex-", or "anti-", or with the second word as in "-là" or "-né". Finally, the hyphen may appear in the text even though it is not associated with any word. The quote can have two different separations: it can be word internal ("aujourd'hui", "o'Donnel", "hors-d'oeuvre"), or may be part of the first word ("l'ami"). Similarly the period may be part of a word, for instance, "L.A.", "sec." (secondes), "p." (page) or may be just a punctuation marker.

Word recognition results using the si-3k and si-38k acoustic training data are summarized in Table 9 for the 5k and 20k vocabularies. Test data consisting of 200 sentences (25 from each of 8 speakers) for each vocabulary were selected from the development test material (Feb94-dev) for a closed vocabulary test. An additional 200 sentences from the development material were used for a 20k open test set [6]. The perplexity of the 5k bigram is 106 and that of the 20k closed bigram is 178. For the 5k test, 428 CD models trained with the si-3k data give a word error of 12.6%. Using 2964 CD models trained on the si-38k data, the word error with the bigram is reduced by 30%. The use of a trigram LM gives an additional 34% reduction of error to 5.7%. For the 20k closed vocabulary test, the si-38k model set gives an error reduction of 29% over the si-3k model set. The use of the trigram LM reduces the word error by an additional 24%. On the 20k open (20k +) test, the word error with the bigram LM is 18.4%. For this data 3.9% of the words are out-of-vocabulary and occur in 72 of the 200 sentences. Comparing the 20k open and closed results, it can be seen that not only are there more substitutions for the open test (10.8% versus 6.8%) the insertion rate is also much higher (3.7% versus 0.9%). Thus apparently the OOV words are not simply replaced by another word, but are more often replaced by a sequence of words. For example, the word "*endeuillé*", which is not in the lexicon, was recognized as the sequence of words "*en deuil et*", which has the same sequence of phonemes. Due to the OOV words the use of a trigram LM only reduces the word error by 16% for the open vocabulary condition.

Table 9
Word recognition results on the Feb94 test data with bigram/trigram grammars estimated on the 38M-word *Le Monde* training text

Conditions	Corr.	Subs.	Del.	Ins.	Err.
428m, si-3k, 5k, bg	88.7	7.5	3.7	1.4	**12.6**
2964m, si-38k, 5k, bg	92.4	5.7	1.9	1.1	**8.7**
2964m, si-38k, 5k, tg	95.3	3.6	1.1	1.0	**5.7**
428m, si-3k, 20k, bg	85.5	11.9	2.6	1.8	**16.3**
2964m, si-38k, 20k, bg	89.9	8.7	1.4	1.4	**11.5**
2964m, si-38k, 20k, tg	92.2	6.8	1.0	0.9	**8.7**
2964m, si-38k, 20k+, bg	86.1	12.7	1.2	4.4	**18.4**
2964m, si-38k, 20k+, tg	88.1	10.8	1.1	3.7	**15.6**

[a] 5k: 5000 word lexicon, 20k: 20,000 word lexicon, 20k+: 20,000 word lexicon with open test.

6. Discussion and summary

In this section we present some observations made on these experiments in large vocabulary, speaker-independent continuous speech dictation using the WSJ and BREF corpora. While we have attempted to define comparable experimental conditions for English and French, there are nonetheless several important differences that should be highlighted. One uncontrollable source of variability is that the test data necessarily come from different speakers. Other differences are in the preprocessing of the text materials and the treatment of case, and in the definition of the

[6] The closed/open vocabulary distinction made for BREF is different from that of WSJ. For BREF, the open and closed conditions share the same lexicon and language model, but the test data is different. For WSJ, the test data is fixed for both conditions, but the lexicon and language model differ.

Table 10
Phone and word error rates with bigram LM for WSJ

Condition	Feb92-5k-si-nvp phone error	Nov92-5k-si-nvp word error
493 models, Δ	30.1	9.7
493 models, ΔΔ	27.6	8.3
884 models, ΔΔ	25.9	6.9
3306 models, ΔΔ	19.2	4.8

open and closed vocabulary tests. For issues specifically related to speech recognition in French, see (Gauvain et al., 1994a). In Section 6.1 we point out the role of phone recognition in system development. Next we attempt to address the issue of why, even though phone recognition accuracy is higher for French than for English, word recognition is better for English than for French. Finally, we point out some common problems discovered in our analysis of the recognition errors.

6.1. Links between phone and word recognition

Much of our development work has relied on the use of phone recognition in order to improve the acoustic models. Evaluating phone recognition enables us to assess the quality of the acoustic models without lexical or higher order constraints. It is also easier and faster to test out ideas using phone recognition than using word recognition. In Table 10 phone error rates on development data from WSJ0 (Feb92-5k-si-nvp) and corresponding word error rates on the Nov92 5k-nvp test data are given. Improvements in speaker-independent phone accuracy on the development data are seen to yield improvements in word accuracy on independent test data. While the same trends in phone recognition are observed on the Nov92-5k-si-nvp, the demonstration on independent data provides direct evidence that it is worthwhile to run phone recognition experiments to measure improvements in acoustic modeling. In addition, comparing the outputs of the phone recognizer and the word recognizer on the development data has led to improvements in the lexical pronunciations and phonological rules.

Similar results were observed for BREF, as shown in Table 11. For both the 5k and 20k closed test data, increasing the number of CD models reduces the phone and word errors. In general, when a large reduction in phone error is obtained between two model sets, the word error is also reduced, both on average and for all of the speakers.

Table 11
Phone and word error rates for BREF with a bigram LM on the Feb94-dev 5k and 20k-closed test

Model set	BREF-5k		BREF-20k	
	Phone error	Word error	Phone error	Word error
428 models	20.5	12.6	20.0	16.3
1747 models	14.4	9.0	14.4	12.9
2964 models	13.5	8.7	13.4	11.5

6.2. Language-dependent differences in word recognition

Even though better phone recognition accuracies are obtained for BREF than for WSJ, word recognition in English is better than in French. This may be in part due to the higher lexical ambiguity for French. To allow comparison of lexical ambiguity for French and English, Table 12 gives homophone rates found in both the training lexicon and texts of BREF and WSJ, where homophone rate is defined to be the number of words which are homophones (words having the same pronunciation), divided by the total number of words. 35% of the words in the 10,311-word BREF si-3k training lexicon are homophones, compared to 6% in 8996-word WSJ0 training lexicon. In the WSJ training texts, 1 out of 5 words is ambiguous, given a perfect phonemic transcription. For BREF, over half the words in the training text have an ambiguous phonemic

Table 12
Left: Single word homophones in BREF and WSJ. Right: Table entries correspond to the number of homophone classes with k graphemic forms in the class

Corpus	Homophone rate		Homophone class size (k)			
	Lexicon	Text	1	2	3	≥ 4
BREF	35%	57%	6686	1329	215	73
WSJ	6%	18%	8453	237	22	1

transcription. In the right part of Table 12 is shown the number of homophone classes of size k, where a homophone class is a set of graphemic words with the same phonemic transcription. For the WSJ0 lexicon, the largest homophone class 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*. While it is difficult to estimate the frequencies of multiple-word and multi-word homophones, we have observed on test data that these too are more frequent in French than in English.

Not only does one phonemic form correspond to different orthographic forms, the reciprocal situation is also rather common. That is, there can be a relatively large number of possible pronunciations for a given word. In English many of the alternate pronunciations are word-internal differences in vowel color, or are due to the reduction of unstressed syllables in polysyllabic words. In the WSJ training lexicon, about 10% of the entries have multiple pronunciations. In the expanded BREF training lexicon about 40% of the entries have multiple pronunciations. This is mainly due to optional word-final phones, such as an optional mute-e insertion for all words ending in a final consonant, and to *liaison* consonants and the optional reduction of word-final consonant clusters. For example the word "autres" can have 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. Using probabilities for each transcription can be useful, but their automatic training is not straightforward and requires a lot of data.

Another problem that has not yet been pointed out, is that for French, a bigram LM is less effective than for English. This is due to the high lexical ambiguity in French—a proper graphemic transcription with error-free agreement in gender and number is unlikely with short term LMs like bigrams. Moreover, correct agreement in gender and number can sometimes be carried out only by relying on semantic knowledge, as neither the acoustic nor the lexical information is sufficient.

Table 13
Comparison of lexical coverage for WSJ and BREF. The numbers in parentheses include case distinctions

	WSJ	*Le Monde*
Text size	37.2M	37.7M
Number of words	165k	259k (280)
5k	90.6%	85.5% (85.2)
10k	94.9%	90.9% (90.6)
20k	97.5%	94.9% (94.6)
40k	99.0%	97.6% (97.3)
80k	99.7%	99.0% (98.9)

A related issue has to do with lexical coverage for a given size lexicon. On average, the lexical coverage for French is less than that for the same size lexicon of English. This is clearly demonstrated in Table 13. We see that the word coverage for BREF with a 5k lexicon is 85%, compared to 91% for WSJ. Similarly, the 20k BREF lexicon has a word coverage of 95% which is significantly smaller than the coverage of the 20k WSJ (98%). For easier comparison with WSJ, the BREF counts were computed without distinguishing case. The numbers in parentheses give the coverage if case distinctions are made. It appears that the lexicon size for French must be doubled to obtain the same coverage as in English.

6.3. Common problems to English and French

We have observed for both English and French that the phone and word recognition accuracies

Table 14
Phone and word error rates for WSJ and BREF speakers on 5k nvp test with 38k training utterances and trigram LM

WSJ – Nov92-5k nvp				BREF – Feb94-5k			
ID	Sex	Phone error	Word error	ID	Sex	Phone error	Word error
440	M	15.1	2.0	IL	M	7.8	3.0
441	F	24.9	7.4	IM	M	14.3	3.9
442	F	15.5	3.3	IN	F	11.2	3.8
443	M	15.5	1.1	IO	M	17.2	6.7
444	F	16.5	3.1	IP	F	14.7	10.5
445	F	17.5	1.7	IR	F	13.0	4.1
446	M	11.7	1.6	IS	M	12.8	4.9
447	M	19.6	4.4	IT	M	16.1	7.6
Average		16.9	3.1	Average		13.5	5.7

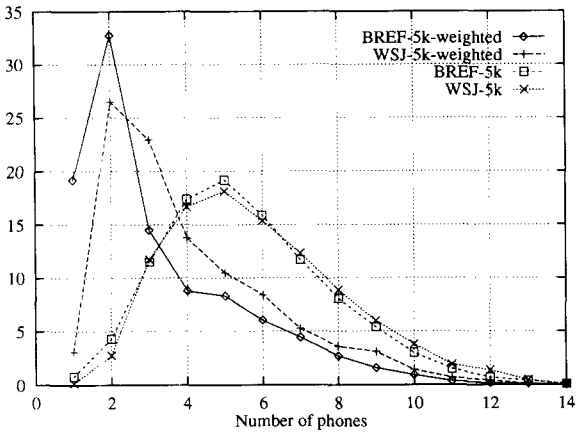

Fig. 3. Word distribution in 37M-word WSJ and 4M-word *Le Monde* texts for the 5k lexicons as a function of the word length in phones.

can differ quite a bit across speakers. This is of course nothing new, as speakers are commonly classified as "sheep" or "goats". To give an idea of the variability across speakers, phone and word errors on the 5k test data are given for WSJ and BREF in Table 14. These results are for the best configuration (i.e. the largest model set and trigram LM) for each language. As can be seen there is a range in performance for both phone and word error. The word error ranges from 1.1% to 7.4% for WSJ and from 3.0% to 10.5% for BREF. Efforts must be taken to determine why the poor speakers are hard to recognize. This may be for a variety of reasons, from the acoustics of their speech production, to their choice of word pronunciations and phonological variants, or their speaking rate.

A large number of errors for both languages involve short words of one or two phonemes. While there are relatively few of these words, they are very frequent, accounting for about 50% and 30% of all word occurrences in French and English respectively. Fig. 3 shows the distribution of words in the 5k lexicons for BREF and WSJ, as a function of the word length in phones. The curves labeled "weighted" reflect the word occurrences in the training text materials. While the distributions in the lexicons are seen to be quite similar, there is a large disparity in the number of monophone words in the running text (almost 20% for *Le Monde* compared to 3% for *WSJ*).

In the 1000 most frequent French words there are 30 monophone words transcribed by 17 different phones. These include almost all of the vowels (except the schwa vowel /x/, and the two open vowels /X/ as in "*leur*" and /c/ as in "*botte*"). The monophone consonant words are all due to the apostrophe. For comparison, in the WSJ 5k lexicon there are 8 monophone words (all vowels, four of which are frequent words "*a*", "*I* ", and reduced forms of "*are*" and "*or*").

This makes French word recognition particularly difficult, as about 20% of the running text are acoustically highly variable monophone words with no intraword phonotactic constraints and with low LM costs (as they are also very frequent). We have observed that nearly any word sequence can be transcribed by a larger number of short, frequent words, resulting in multiword homophones. Some examples of recognition errors where the longer word has been split in a sequence of shorter words, with no or minor errors in the phonetic transcription are "*désengagement → des engagements*", "*couteaux → coûts taux*", and "*il laisse → il et se*". Errors on monophone words account for 20% of the substitutions, 75% of the insertions, and 85% of the deletions. Two-phone words account for an additional 30% of the substitutions and essentially all the remaining insertions and deletions.

The larger number of monophones in French also contributes to the increase in word error rate when there are OOV words. In English, we have observed roughly 1.2 errors for each OOV word on the 20k + test. In French, not only are there more OOV words for the 20k + test (3.9% compared to 2.0% for English), the word error increases by 6.9%, indicating that most OOV words are replaced by a sequence of words.

In English, short words, mostly function words, account for about 80% of the deletions, 80% of the insertions, and 45% of the substitutions. Some typical errors involve inflected forms of verbs such as "*finishing → finish in*" or "*expect it → expected*". These are almost multiword homophones. In the first case the error seems to arrive from acoustic causes, where the "*ing*" is recognized as "*in*" (the signal was listened to to verify that the speaker actually did say "*ing*"), and in

the second case the cause is the language model in that "*expect it*" has a higher probability than "*expected*".

In French, another major source of error with a bigram LM involves the insertion or deletion of mute-e, or a monophone word that is the same as one of the surrounding phonemes. For example, the word sequence "*en priorité un*" was misrecognized "*en priorité et un*". This kind of error is difficult to handle as on the acoustic level it requires refined duration models, and on the LM level a longer span model than a bigram is needed. Deletion problems also involve mostly monophone words, where the reasons for deletion are similar to those for insertion. If all the words in the lexicon have an optional final mute-e, assimilation of an adjacent monophone word may result in increased deletions. In contrast, if the lexicon does not allow a mute-e at the end of a word, the system will insert a short word if the mute-e is pronounced.

Another problem with using a bigram-backoff LM is that the most frequent words (in particular the monophone words, as shown in Fig. 3) have the highest backoff LM scores and thus appear easily in place of acoustically similar words which had fewer observations in the training text. This problem was observed for long words with low counts in the training corpus: they were often recognized as a sequence of small words with identical phonemic transcriptions. Many of the homophones in French which arise from different gender or number forms may also be insufficiently handled by the bigram-backoff LM. If the bigram does not exist and the backoff component is used, the more frequent form will be chosen without regard to agreement.

The use of a trigram LM was found to correct some agreement errors, in gender, number and negation. In French a negative form is usually made by surrounding the verb with "*ne* VERB *pas*". While with the bigram the "*ne*" can be easily deleted, the trigram is able to capture this constraint. The use of a trigram LM was shown to improve the recognition accuracy by 20% to 30% over a bigram LM. The use of N-class language models (as opposed to N-grams) can be helpful for French, where the number of different graphemic forms for a given root form is much higher than for English.

6.4. Summary

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 of phonological variations, and decoding. We have described our system and an evaluation on two dictation tasks using read, newspaper-based corpora: the ARPA *Wall Street Journal* corpus of American English and the BREF *Le Monde* corpus of French; and on the ARPA Resource Management task that has been widely used to evaluate and compare systems. The decoder uses a time-synchronous graph-search strategy for a first pass with a bigram backoff language model, which includes intra- and interword context-dependent phone models, intra- and interword phonological rules, phone duration models, gender-dependent models. When a trigram LM is used, a second acoustic decoding pass is carried out using the word graph generated in the first pass.

High precision acoustic modeling is achieved with continuous density HMMs and large amounts of training data, with which phone accuracies on the order of 81% for English and 87% for French are obtained. Word recognition experiments were presented for BREF and WSJ using 5k and 20k vocabularies with bigram and trigram language models. The use of a trigram LM in a second pass gave an error reduction of about 30% for a closed vocabulary test and about 15% for an open vocabulary test relative to the bigram results. Word accuracies of 96.9% on WSJ and 94.3% on BREF have been obtained for a 5000 word vocabulary. With 20,000 word lexicons and an open vocabulary test the word accuracy is on the order of 90% for WSJ and 85% for BREF. This difference in word error can be attributed to problems such as the larger number of out-of-vocabulary words in French (an effect of the lower word coverage for the lexicon), the higher number of homophones and monophone words, liaison, mute-e, and gender and number agreement.

Acknowledgments

The authors express their thanks to Murray Spiegel (Bellcore) for providing ORATOR phonetisizations for a subset of the WSJ lexicon, and to the anonymous reviewers for providing valuable comments on the original manuscript.

References

M. Cohen (1989), Phonological Structures for Speech Recognition. PhD Thesis, University of California, Berkeley., CA.

S.B. Davis and P. Mermelstein (1980), "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 28, No. 4.

J.L. Gauvain and L.F. Lamel (1992), "Speaker-independent phone recognition using BREF", *Proc. DARPA Speech and Natural Language Workshop*, February 1992.

J.L. Gauvain and C.H. Lee (1992), "Bayesian learning for hidden Markov model with Gaussian mixture state observation densities", *Speech Communication.*, Vol. 11, Nos. 2–3, pp. 205–213.

J.L. Gauvain and C.H. Lee (1994), "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE Trans. Speech Audio Process.*, Vol. 2, No. 2, April.

J.L. Gauvain, L.F. Lamel and M. Eskénazi (1990), "Design considerations and text selection for BREF, a large French read-speech corpus", *Proc. ICSLP-90*.

J.L. Gauvain, L.F. Lamel, G. Adda and J. Mariani (1994a), "Speech-to-text conversion in French", *Internat. J. Pattern Recogn. Artif. Intell.*, Vol. 8, No. 1, 1994.

J.L. Gauvain, L.F. Lamel, G. Adda and M. Adda-Decker (1994b) "The LIMSI continuous speech dictation system: Evaluation on the ARPA Wall Street Journal task", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.-94.*

E. Giachin, A.E. Rosenberg and C.H. Lee (1991), "Word juncture modeling using phonological rules for HMM-based continuous speech recognition", *Comput. Speech Language*, Vol. 5.

S.M. Katz (1987), "Estimation of probabilities from sparse data for the language model component of a speech recognizer", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 35, No. 3.

L.F. Lamel and J.L. Gauvain (1992), "Continuous speech recognition at LIMSI", *Proc. Final Review of the DARPA ANNT Speech Program*, September.

L.F. Lamel and J.L. Gauvain (1993a), "Cross-lingual experiments with phone recognition", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Pocess.-93.*

L.F. Lamel and J.L. Gauvain (1993b), "High performance speaker-independent phone recognition using CDHMM", *Proc. Eurospeech-93.*

L. Lamel and J.L. Gauvain (1993c), "Identifying non-linguistic speech features", *Proc. Eurospeech-93.*

L.F. Lamel, J.L. Gauvain and M. Eskénazi (1991), "BREF, a large vocabulary spoken corpus for French", *Proc. Eurospeech-91.*

C.H. Lee, L.R. Rabiner, R. Pieraccini and J.G. Wilpon (1990), "Acoustic modeling for large vocabulary speech recognition", *Comput. Speech Language*, Vol. 4.

H. Ney (1984), "The use of a one-stage dynamic programming algorithm for connected word recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 32, No. 2.

D.S. Pallett and J.G. Fiscus (1992), "Resource management corpus—Continuous speech recognition—September 1992 test set benchmark test results", *Proc. Final Review of the DARPA ANNT Speech Program*, September.

D.S. Pallett, J.G. Fiscus, W.M. Fisher and J.S. Garofolo (1993), "Benchmark tests for the DARPA spoken language program", *Proc. ARPA Human Language Technology Workshop*, March.

D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, B.A. Lund and M.A. Pryzbocki (1994), "1993 benchmark tests for the ARPA spoken language program", *Proc. ARPA Human Language Technology Workshop*, March.

D.B. Paul and J.M. Baker (1992), "The design for the Wall Street Journal-based CSR corpus", *Proc. ICSLP-92.*

P. Price, W.M. Fisher, J. Bernstein and D.S. Pallett (1988), "The DARPA 1000-word resource management database for continuous speech recognition", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.-88.*

B. Prouts (1980), 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, November.

L.R. Rabiner, B.H. Juang, S.E. Levinson and M.M. Sondhi (1985), "Recognition of isolated digits using hidden Markov models with continuous mixture densities", *AT&T Tech. J.*, Vol. 64, No. 6.