Cross-Lingual Experiments with Phone Recognition

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 phone recognition for both French and English. The phone accuracy is assessed on the BREF corpus for French, and on the Wall Street Journal and TIMIT corpora for English. Cross-language differences concerning language properties are presented. It was found that French is easier to recognize at the phone level (the phone error for BREF is 23.6% vs. 30.1% for WSJ), but harder to recognize at the lexical level due to the larger number of homophones. Experiments with signal analysis indicate that a 4kHz signal bandwidth is sufficient for French, whereas 8kHz is needed for English. Phone recognition is a powerful technique for language, sex, and speaker identification. With 2s of speech, the language can be identified with better than 99% accuracy. Sex-identification for BREF and WSJ is error-free. Speaker identification accuracies of 98.2% on TIMIT (462 speakers) and 99.1% on BREF (57 speakers), were obtained with one utterance per speaker, and 100% with 2 utterances.

INTRODUCTION

Our long term goals include the development of speech recognizers that are speaker and vocabulary independent, and can be used in multiple languages for large vocabulary tasks including dictation. It is well-known that the problems of speech-to-text conversion can be different from language to language. In this paper some language-dependent issues are addressed for French and English. These include the number of phonemes in the language, the power of phonotactic constraints provided by phone n-grams, and the lexical ambiguity. Other well-known differences including the role of intonation, lexical stress, allophonic variations are not discussed here. Further details about problems specific to speech-to-text conversion in French can be found in [5].

Attention is focused on phone recognition in these two languages in order to assess the relative difficulties of each without the influence of lexical and syntactic constraints. In attempting to have a fair cross-lingual comparison, similar experimental conditions are used i.e., roughly the same amount of speech data, recorded under similar conditions (8kHz bandwidth, close-talking microphone, read-speech) is used to train the models. The French data come from the BREF corpus[6, 10] and the English data come from the DARPA Wall Street Journal corpus[14]. For comparative purposes, the recognizer is also evaluated on the DARPA TIMIT corpus[3]. The same recognizer is used, and is evaluated using sets of context-dependent phone models, where each model is a left-to-right HMM with Gaussian mixture observation densities. The phone accuracies reported here do not reflect the best performance for each task, but rather attempt to make cross-task comparisons with similar conditions. Acoustic processing is addressed, including bandwidth and acoustic feature choice. The power of phone recognition for non-linguistic speech feature identification is demonstrated.

SPEECH CORPORA

These efforts use large corpora of read speech material from a large number of speakers, with the aim of building base acoustic models which can be augmented and adapted to specific speakers or tasks. By using read-speech, the text materials can be selected so as to control for different events such as the phonetic contexts. The material in BREF was selected to maximize the number of different phonemic contexts, whereas the WSJ texts were selected so as to contain only words in the most frequent 64,000 words in the original text material. A subset of the material in TIMIT was selected to cover rare, yet potentially "interesting" phonemic environments. 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.

The BREF Corpus: BREF is a large read-speech corpus, containing over 100 hours of speech material, from 120 speakers (55m/65f)[10]. 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[6]. Containing 1115 distinct diphones and over 17,500 triphones, BREF can be used to train VI phonetic models. The text material was read without verbalized punctuation.

In these experiments approximately 4.3 h of speech data are used for training. This represents 2770 sentences from 57 speakers (28m/29f). The test data consists of 93 sentences from 8 speakers (4m/4f). The test text material is distinct from the training texts. Phone transcriptions of these utterances were automatically generated and manually verified[4].

The DARPA WSJ Corpus: The DARPA Wall Street Journal-based Continuous-Speech Corpus (WSJ)[14] has been designed to provide general-purpose speech data (primarily, read speech data) with large vocabularies. Text materials were selected to provide training and test data for 5K and 20K word, closed and open vocabularies, and with both verbalized and non-verbalized punctuation. The recorded speech material supports both speaker-dependent and speaker-independent training and evaluation.

In these experiments the standard SI-84 training material, containing 7240 sentences from 84 speakers (42m, 42f) is used to build the phone models. The non-verbalized-punctuation and verbalized-punctuation DARPA Feb92 pilot evaluation test material are used for test. This data consists of 200 sentences from 10 speakers (6m/4f) for each condition. Since there are no associated phone transcriptions for this data, the "correct" phone transcription was determined by performing segmentation allowing multiple pronunciations for words, and optional phonological rules to be applied at

word boundaries.

The DARPA TIMIT Corpus: The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus[3] is a corpus of read speech designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the U.S.

The TIMIT CDROM[3] contains a training/test subdivision of the data that ensures that there is no overlap in the text materials. The subdivision provides 10 sentences from each of 462 speakers for training. In these experiments, the core test set containing 8 sentences from each of 24 speakers (2m/1f from each dialect region) is used for testing. All of the utterances in TIMIT have associated time-aligned phonetic transcriptions.

LANGUAGE CHARACTERISTICS

Phone sets: A set of 35 phones are used to represent the French data. These contain 14 vowels (including 3 nasal vowels), 20 consonants (6 plosives, 6 fricatives, 3 nasals, and 5 semivowels), and silence. The phone table can be found in [4]. For English, a set of 46 phones are used for WSJ and the standard set of 61 phone symbols are used for TIMIT[3]. Different symbol sets are used because TIMIT is transcribed at a broad phonetic level, but for WSJ the phone sequence is obtained by concatenating the phonemes for the lexical entries in the associated text string. In TIMIT, plosives are represented by a sequence of a closure followed by a release, whereas for WSJ they are represented by a single symbol. Other allophones, such as the voiced-h, the fronted-u, the flaps, and glottal stop, found in TIMIT are not used in WSJ. TIMIT also distinguishes 3 types of silence, whereas only one is used in WSJ. English is thus represented using 21 vowels (including 3 diphthongs and 3 schwas), 24 consonants (6 plosives, 8 fricatives, 2 affricates, 3 nasals, 5 semivowels), and silence, plus several finer distinctions for TIMIT.

Phone perplexity: One way to compare the complexity of phone recognition across languages is to look at the phone perplexities. These are given in Table 1 for the training and test corpora for BREF, WSJ, and TIMIT, along with the number of phones and diphones occuring in the training material. Comparing BREF and WSJ, it can be seen that although French has fewer phones, the training perplexities are about the same, and WSJ has a higher test perplexity for the non-verbalized punctuation (nvp) and a lower perplexity for the verbalized punctuation (vp).

For TIMIT the perplexities are given using the 61 phone set and also a reduced 39 phone phone set[12]. It should be noted that the phone perplexity computed on the training material of TIMIT is estimated on a much smaller text set, than those estimated for BREF and WSJ. While the TIMIT training material contains 4620 utterances, there are fewer than 2000 different text prompts, which probably explains the larger difference in training/test perplexity than is observed for BREF and WSJ.

Lexical ambiguity: Even though French has fewer phonemes than English, the lexical ambiguity as measured by homophone rate is much higher for French. The homophone rate is defined to be the number of words which are homophones (having the same pronunciation as another word), divided by the total number of words. Table 2 gives the homophone rates for BREF and WSJ, counted on the lexicon and on the training texts. The latter provides a frequency weighted estimate of the homophone rate. In the 10,311-word BREF training lexicon, 35% of the words are homophones, compared to 6% in 8996-word WSJ training lexicon. In the WSJ training texts, 1 out of 5 words is ambiguous, even given a perfect phonemic transcription. For BREF, over half the words in the training text are ambiguous. The right part of the table gives the number of orthographic words associated with a pronunciation having a given size homophone class. For the WSJ lexicon, the largest homophone class has 4 entries: *B., Bea, bee, and be*. In the BREF lexicon there are 3 pronunciations having 7 orthographic words, as in *100, cent, cents, san, sang, sans, sent.*

Corpus	*BREF*	*WSJ nvp/vp*	*TIMIT-61*	*TIMIT-39*
# phones	35	46	61	39
# diphones	1160	1571	2461	1139
training px.	16.2	16.8	15.7	12.8
test px.	16.1	17.5/15.1	18.9	14.6

Table 1: Phone perplexities computed on the training data.

	Homophone rate		*#prons/#words in set*			
Corpus	*Lexicon*	*Text*	*1*	*2*	*3*	≥ 4
BREF	35%	57%	6686	1329	215	73
WSJ	6%	18%	8453	237	22	1

Table 2: Single word homophones in BREF and WSJ.

EXPERIMENTS IN PHONE RECOGNITION

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. Duration is modeled with a gamma distribution per phone model. As proposed by Rabiner et al.[15], 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[8] and moment estimators for the gamma distributions.

The phone recognition results given here use sets of context-dependent (CD) models which were automatically selected based on their frequencies in the training data. There are 428 models for BREF, 488 for WSJ, and 459 for TIMIT. While we have obtained higher performances using more models, there is a substantial increase in computation.

The overall Markov chain is obtained by connecting the phone HMMs 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 for BREF, the resulting HMM includes 1294 non-null states and has about 1,070,00 parameters.

Experiments were run varying the signal analysis used to compute the cepstrum coefficients (LPC or Fourier analysis), and for two bandwidths (4kHz and 8kHz). In all cases, a 30 ms window was used with a 10 ms frame rate. For the 4kHz bandwidth, the 16kHz speech was downsampled by 2, and a 26-dimensional feature vector composed of LPC-derived cepstrum coefficients and the first-order time derivative is used. It was found that no significant difference was observed by using LPC or DFT based cepstra with a 4kHz bandwidth[9]. For the 8kHz Fourier analysis, a 32-component feature vector consisting of 16 Bark-frequency scale cepstrum coefficients and their first order differences, is computed on the 8kHz bandwidth. For each frame, 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]. For 8kHz bandwidth, this analysis was found to outperform an LPC-based analysis even with frequency warping.

Model set	*BREF*	*WSJ (nvp/vp)*	*TIMIT (61/39)*
4k LPCC	23.8	33.9/29.1	39.3/32.8
8k MFCC	23.6	30.1/26.9	37.2/30.9

Table 3: Phone error with CD models with phone bigram.

The phone errors rates are given in Table 3.[1] The 8kHz MFCC analysis consistently improves performance for English (on the order of 2%), but there is almost no improvement for French. The 8kHz phone error rate for BREF is 23.6%, compared to 30.1% for WSJ nvp. As expected, the error for vp is lower, 26.9%. The phone error on TIMIT is 37.2% with the 61 phone set, and 30.9% when mapped to a 39 phone set[12]. Better results were obtained with the same model set by including 2nd order derivatives, 34.9% (61 phones) and 28.9% (39 phones). To our knowledge, the best reported results on the core test are by Robinson[16]: 31.3% (61 phones) and 26.1% (39 phones).

It is interesting to note that higher phone accuracies are obtained for BREF, even though the phone perplexity is about the same as WSJ. It may be simply that the phonetic structure of French is easier to recognize than that of English. French has fewer consonant clusters than English, and has a more regular consonant-vowel alternation. French vowels are acoustically relatively stable compared to American English ones whose spectral characteristics vary more within the segment.

This may be related to the differences in lexical ambiguity. Perhaps the large number of homophones in French require clearer acoustics, since phonetic recognition errors will largely increase the number of word candidates. Since in English there are fewer homophones, there may be more freedom in the phonetic realization.

Error analysis: Table 4 summarizes the most common substitution errors for BREF and WSJ nvp. Substitutions account for 15.2% of the errors on BREF and 16.6% of the errors on WSJ. In French the most common confusions are among the vowels, with symmetric /e,E/ and /E,a/ confusions being the most frequent. The confusability between /e,E/ arises because in some word positions, this distinction is not necessary for unambiguous interpretation. For English the most common substitutions are between the vowels /I,x/ and in voicing for /s,z/ and /t,d/. The vowel confusions are probably in part due to their somewhat arbitrary specification in the lexicon, as well as to insufficent duration models, which may also contribute to the consonant voicing errors. The errors are also likely to be related to stress and syllable position, which are not included in the contexts.

[1]For BREF and WSJ phone errors are reported after removing silences, whereas for TIMIT silences are included as transcribed. Scoring without the sentence initial/final silence increases the phone error by about 1.5%.

BREF		*WSJ nvp*	
e → E	5.9%	I → x	3.1%
E → a	3.7%	z → s	2.8%
E → e	3.3%	x → I	2.7%
I → a	3.1%	d → t	2.3%
a → E	2.0%	t → d	1.8%

Table 4: The most common substitution errors.

Deletions account for 4.2% and insertions for 3.5% of the errors on BREF. The phones /r,l/ account for over 20% of the insertions and deletions, which is logical as they have the shortest average duration. For WSJ, there are 6.5% deletions and 4.9% insertions, with /t,x,d,n/ having the most deletions and /t,d,x/ accounting for 36% of the insertions. These phones can also be very short.

LANGUAGE IDENTIFICATION

An 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 normalized likelihood. Experiments were reported using sets of SI CI phone models for French and for English[9], with a 4kHz LPCC analysis. Using this approach and processing the entire utterance always gave 100% correct language identification for taken from 8 speakers (4m/4f) of each language.

This technique has been further investigated using more test data. For English, the 120 SX sentences in the TIMIT coretest are used, and for French, a set of 130 BREF sentences from 21 speakers (10m,11f) are used. SI, CI models are used with out a phone bigram so as to minimize incorporation of task information. All of the sentences were adjusted to have only 100 ms of silence at the beginning/end. The numbers of errors for each language are given as a function of duration in Table 5. With as little as 400 ms of signal, there is less than 3% error in language identification, and with 1.2s, language identification is error free. To test the task dependence of this technique, language identification on 100 sentences from WSJ nvp was evaluated using the same models and system parameters. In this case they are more misclassifications, however with 2s of speech the accuracy is 97% (with 4s they are no errors). These results show that task independent language identification is feasible using this approach.

Duration	*#sents*	*0.4s*	*0.8s*	*1.2s*	*1.6s*	*2.0s*
French	130	2	3	1	0	0
English TIMIT	120	4	1	0	0	0
English WSJ	100	15	8	7	4	3

Table 5: Language identification as a function of duration and language.

IDENTIFICATION OF OTHER 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

and the identity of the speaker.

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 used in our 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 on the Feb92 or Nov92 5K test data. Sex identification on the 192 sentences in the TIMIT core test set resulted in one error on a short sentence from a male speaker. However, the phone accuracy on this sentence was higher using the female SI models than with the male SI model set. For BREF, no errors were observed on the 93 test sentences from the 10 test speakers, nor on an additional 16 sentences from 9 speakers.

Speaker Identification: For speaker identification, a set of CI phone models were built for each speaker, by supervised adaptation of SI models[7]. Since TIMIT contains speech from a large number of speakers, and has recently been used for speaker identification[1, 13, 17], it was decided to use this corpus for evaluation in English. The reported results have shown high speaker identification rates using subsets of 100 to all 462 speakers, indicating that speaker-identification on this data should be relatively easy. A speaker-independent set of 30 CI models were built using data from all of the 462 training speakers. These models were then adapted to each speaker using 8 sentences (2 SA, 3 SX, and 3 SI). The remaining 2 SX sentences for each speaker were reserved for the identification test. While the original CI models had a maximum of 32 Gaussian mixtures, the adapted models were limited to 4 mixture components, since the amount of adaptation data was relatively limited.

The unknown speech was recognized by all of the speakers models in parallel. Experiments for English using all 462 speaker models in parallel resulted in 98.2% correct identification using 1 sentence for identification and 100% if both sentences were used. With one sentence, the identification rate for the 136 female speakers was 98.9%, compared to 97.9% for the 326 male speakers.

For French, the base acoustic models were the 35 CI BREF models, built using the training data from the 57 training speakers. In order to have a similar situation to English, these models were adapted to each speaker using only 8 of the training sentences, and 2 sentences for identification test. Using only one sentence per speaker for identification, there is one error, giving an identification accuracy of 99.1%. As for TIMIT, when 2 sentences are used all speakers are correctly identified.

When there was a confusion, the speaker was always identified by another speaker of the same sex. Thus, a simple reduction in computation can be gained by first determining the sex of the speaker by running in parallel SI male and female models. Further reductions in the computation required during recognition can be obtained by speaker clustering.

SUMMARY

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. In this paper, phone recognition performance is compared for English and French on similar corpora. French is easier to recognize at the phone level (the phone accuracy for BREF is 76.4% vs. 69.9% for WSJ), but harder to recognize at the lexical level due to the larger number of homophones. Experiments with signal analysis indicate that a 4kHz signal bandwidth is sufficient for French, whereas 8kHz is needed for English. Phone recognition is shown to be a powerful technique for language identification. With 2s of speech the language is correctly identified as English or French with 99% accuracy. Phone recognition also gives accurate sex and speaker identification. Sex-identification for BREF and WSJ was error-free, and over 99% accurate for TIMIT. Speaker identification accuracies of 98.2% on TIMIT (462 speakers) and 99.1% on BREF (57 speakers), were obtained with one utterance per speaker, and 100% if 2 utterances are used for identification.

ACKNOWLEDGEMENT

The authors express their thanks to Murray Spiegel (Bellcore) for providing ORATOR phonetisizations for a subset of the WSJ lexicon.

REFERENCES

[1] Y. Bennani, "Speaker Identification through a Modular Connectionist Architecture: Evaluation on the TIMIT Database," *ICSLP-92*.

[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, L.F. Lamel, "Speaker-Independent Phone Recognition Using BREF," *DARPA Speech & Nat. Lang. Workshop*, Feb-92.

[5] J.L. Gauvain, L.F. Lamel, G. Adda, J. Mariani, "Speech-to-Text Conversion in French," to appear in *Int. J. Pat. Rec. & A.I.*, 1993.

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

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

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

[9] L.F. Lamel, J.L. Gauvain, "Continuous Speech Recognition at LIMSI," *DARPA Continuous Speech Recognition Workshop*, Sep-92.

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

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

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

[13] C. Montacié, J.L. Le Floch, "AR-Vector Models for Free-Text Speaker Recognition," *ICSLP-92*.

[14] D. Paul, J. Baker, "The Design for the Wall Street Journal-based CSR Corpus" *DARPA Speech & Nat. Lang. Workshop*, Feb-92.

[15] 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.

[16] T. Robinson, "Several improvements to a recurrent error propagation phone recognition system," *Tech. Rep. CUED/TINFENG/TR.82*, 1991.

[17] M. Savic, J. Sorenson, "Phoneme Based Speaker Verification," *ICASSP-92*.