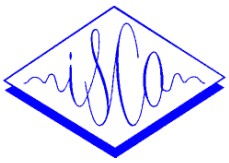

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

ISSUES IN LARGE VOCABULARY, MULTILINGUAL SPEECH RECOGNITION*

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

LIMSI - CNRS, B.P. 133, 91403 Orsay, France
{lamel,madda,gauvain}@limsi.fr

ABSTRACT

In this paper we report on our activities in multilingual, speaker-independent, large vocabulary continuous speech recognition. The multilingual aspect of this work is of particular importance in Europe, where each country has its own national language. Our existing recognizer for American English and French, has been ported to British English and German. It has been assessed in the context of the LRE SQALE project whose objective was to experiment with installing in Europe a multilingual evaluation paradigm for the assessment of large vocabulary, continuous speech recognition systems. The recognizer makes use of phone-based continuous density HMM for acoustic modeling and *n*-gram statistics estimated on newspaper texts for language modeling. The system has been evaluated on a dictation task with read, newspaper-based corpora, the ARPA *Wall Street Journal* corpus of American English, the WSJCAM0 corpus of British English, the BREF-*Le Monde* corpus of French and the PHONDAT-*Frankfurter Rundschau* corpus of German. Under closely matched conditions, the average word accuracy across all 4 languages is 85%, obtained with an open-vocabulary test and 20k trigram systems (64k system German).

10.21437/Eurospeech.1995-46

INTRODUCTION

Speech recognition research at LIMSI aims to develop recognizers that are task-, speaker-, and vocabulary-independent so as to be easily adapted to a variety of applications for different languages. The applicability of speech recognition techniques used for one language to other languages is of particular importance in Europe. The multilingual aspects of this work are in part carried out in the context of the LRE SQALE (Speech recognizer Quality Assessment for Linguistic Engineering) project, which aimed at assessing language-dependent issues in multilingual recognizer evaluation[11]. In the SQALE project, the same system is being evaluated on comparable tasks in different languages (American English, British English, French and German) to determine cross-lingual differences, as well as different systems on the same data so as to compare different methods. A baseline condition for comparison of systems using the same acoustic training data, the same vocabulary and the same language model has been defined for each language. This contribution addresses issues in using a given recognition technology for different languages in the framework of large vocabulary, speaker-independent, continuous speech recognition and discuss language-specific observations.

The recognizer makes use of continuous density HMM with Gaussian mixture for acoustic modeling of context-dependent phone models. For language modeling *n*-gram statistics are estimated on text material. To deal with phonological variability alternate pronunciations are included in the lexicon. In order to limit the decoder search space, a progressive decoding technique is used to reduce the number of hypothesis needed to be evaluated at any given step[2]. In this multiple pass approach larger and more accurate acoustic and language models are used in later passes, with information transmitted via word graphs.

When porting a recognizer to a new language, certain system parameters or components will have to be changed, i.e. those incorporating language-dependent knowledge sources such as the selection of the phone set, the recognition lexicon (alternate word pronunciations), and phonological rules. Other language dependent factors are related to the acoustic confusability of the words in the language (such as homophone, monophone, and compound word rates) and the word coverage of a given size recognition vocabulary. There are other parameters which can be considered language independent, such as the language model weight and word or phone insertion penalties. The selection of these parameters can vary however depending on factors such as the expected out-of-vocabulary rate.

In the remainder of this paper we discuss the language dependencies of the main system components, describe the corpora involved, and for each language the most important characteristics in relation with the recognition technology. Experimental results are provided for all four languages, where acoustic models have been estimated from speech corpora ranging from 7k to 15k utterances and more than 37M words of newspaper text are used for LM training.

LANGUAGE MODELING AND LEXICON

Language modeling entails incorporating constraints on the allowable sequences of words which form a sentence. Statistical *n*-gram models attempt to capture the syntactic and semantic constraints by estimating the frequencies of sequences of *n* words. A backoff mechanism[7] is used to smooth the estimates of the probabilities of rare *n*-grams by relying on a lower order *n*-gram when there is insufficient training data, and to provide a means of modeling unobserved *n*-grams.

The LM training data consists of about 37M words of newspaper text for each language (*WSJ* for English, *Le Monde* for French, *Frankfurter Rundschau*[1] (*FR*) for

*This work is partially funded by the LRE project 62-058 SQALE.

[1]The *Frankfurter Rundschau* LM training texts were obtained from the ACL-ECI CDROM distributed by Elsnet and LDC.

German). In order to be able to construct LMs for the different languages, it was necessary to carry out some kind of language-dependent normalization in order to define a word. Whereas the English texts have all been completely capitalized [9], capitalization is kept as distinctive feature in French and German. For the German texts, the preprocessing was carried out by Philips, and no further processing was done at LIMSI. The French *Le Monde* texts [1] have been normalized and semi-automatically checked for capitalization errors. Table 1 compares some characteristics of these text corpora. In the same size training texts, there are almost 70% more distinct words for *Le Monde* and about 300% more distinct words for *FR* than for *WSJ*. As a consequence, the lexical coverage for a given size lexicon is smallest for *FR* and highest for *WSJ*. For example, the 20k *WSJ* lexicon accounts for 97.5% of word occurrences, but the 20k *FR* lexicon only covers 90.0% of word occurrences in the training texts. Comparing French and English we may observe that for lexicons in the range of 5k to 40k words, the number of words must be doubled for *Le Monde* in order to obtain the same word coverage as for *WSJ*. The difference in lexical coverage for French and English mainly stems from the number and gender agreement in French for nouns, adjectives and past participles, and the high number of different verbal forms for a given verb (about 40 forms in French as opposed to at most 5 in English). German is also a highly inflected language, and one can observe the same phenomena as in French. In addition German has case declension for articles, adjectives and nouns. The four cases: nominative, dative, genitive and accusative can generate different forms for each case which often are acoustically close. For example, while in English there is only one form for the definite article *the*, in German number and gender are distinguished, giving the singular forms *der, die, das* (male, female, neuter) and the plural form *die*. Declension case distinction adds 3 additional forms *des, dem, den* to the nominative form *der*. In German all nouns or substantives are written with capitalized first letters and most words can be substantivized, thus generating lexical variability and homophones in recognition. But the major reason of the poor lexical coverage in German certainly arises from word compounding. Whereas compound words or concepts in English are typically formed by a sequence of words (e.g.: *speech recognition, the speech recognition problem*) or in French by adding a preposition (e.g.: *reconnaissance de la parole, le problème de la reconnaissance de la parole*), in German words are put together to form a new single word (e.g.: *Spracherkennung, das Spracherkennungsproblem*) which in turn include all number, gender and declension agreement variations.

The OOV problem could be reduced in German by a text preprocessing aimed at separating compound words into their constituent building blocks. This step is far from being straightforward and requires a refined morphological analysis. (Decompounding can also result in lower semantic resolution). During the text normalization, the numbers were decompounded in order to increase lexical coverage. Thus, the number 1991 which is written in standard German *neunzehnhunderteinundneunzig* has been changed into the word sequence *neunzehn hundert ein und neunzig*. The backoff technique used to smooth the LM *n*-grams is particularly useful in case of high lexical variability, in order to estimate frequencies of unobserved *n*-grams.

Looking at language-dependent features in lexica and texts, we can observe that the number of homophones is higher for French and German than for English. A comparative study of French and English showed that, given a perfect phonemic transcription, 23% of words in the *WSJ* training texts are ambiguous, whereas 75% of the words in the *Le Monde* training texts have an ambiguous phonemic transcription[3]. In German homophones arise from case-sensitivity and from compound words being recognized as sequences of component words. A major difficulty in French comes from the high number of monophone words. Most phonemes can correspond to one or more graphemic forms (e.g. the phoneme /ɛ/ can stand for *ai, aie, aies, ait, aient, hais, hait, haie, haies, es, est* and /s/ can stand for *s', c'*). The other languages have fewer monophones, and these monophones are considerably less frequent in the texts. Counting monophone words in *Le Monde* and *WSJ* training texts, gave about 17% for French versus 3% for English[3]. In French, not only is there the frequent homophone problem where one phonemic form corresponds to different orthographic forms, there can also be a relatively large number of possible pronunciations for a given word. The alternate pronunciations arise mainly from optional word-final phones, due to liaison, mute-e, and optional word-final consonant cluster reduction.

A common vocabulary list was specified for each language by the SQALE consortium. For American and British English a common 20k word vocabulary (corresponding to the 1993 ARPA WSJ baseline test) was used. The French vocabulary contains the 20k most frequent words in the training text corpus. Due to the lower lexical coverage of German, a 64k-word vocabulary was chosen. The pronunciation lexicons for French and American English were developed at LIMSI. For British English we combined portions of the BEEP dictionary from Cambridge University with pronunciations taken from LIMSI American English WSJ lexicons, which were remapped to be more "British". The source German lexicon supplied by Philips has been progressively modified at LIMSI. Minor changes were made prior to the SQALE dry-run test (Feb95), and more global modification, including a reduction of the phone set were made prior to the eval test.

ACOUSTIC MODELING

The recognizer makes use of continuous density HMM (CDHMM) with Gaussian mixture for acoustic modeling. The acoustic models are sets of context-dependent (CD), position-independent phone models, which include both intra-word and cross-word contexts, selected automatically based on their frequencies in the training data. Each phone model is a 3-state left-to-right CDHMM with Gaussian mixture observation densities (typically 32 components).

The acoustic models are built in a series of steps. A first set of models is used to segment and label the train-

Corpus	*WSJ*	*WSJCAM0*	*Le Monde*	*FR*
language	*American*	*British*	*French*	*German*
Training text size	37.2M	37.2M	37.7M	36M
#distinct words	165k	165k	280k	650k
5k coverage	90.6%	90.6%	85.2%	82.9%
20k coverage	97.5%	97.5%	94.7%	90.0%
20k-OOV rate	2.5%	2.5%	5.3%	10.0%

Table 1: Comparison of *WSJ*, *WSJCAM0*, *Le Monde*, and *Frankfurter Rundschau* lexica and LM training corpora.

ing data using Viterbi alignment of the text transcription and a lexicon containing one or more pronunciations per word. The chosen phone sequence and segmentation are then used to construct a set of context-independent models, with a maximum of 32 Gaussians per state. The training data is then resegmented using these models, and context-dependent model sets are built.

Table 2 compares some characteristics of the four speech corpora being used for this work. The speech data come from the ARPA WSJ corpus for American English[9], the WSJCAM0 corpus for British English[10], BREF for French[8], and Phondat for German[2]. Phondat is not read newspaper speech, but contains phonetically balanced sentences, some short stories, isolated letters and train timetable queries. In table 2 we show the number of phones used for each language: for French we use only 34 phone symbols, whereas for German we started the development of the system with a set of 51 symbols (ignoring glottal stop), the size of which has been reduced after the dryrun test to 48 (including glottal stop). The reduction was obtained by ignoring the *long-vowel* distinction for a subset of vowels. The glottal stop model is specific to the german system. This symbol has no distinctive role concerning the phonetic transcription of an isolated word, but in continuous speech its presence indicates a word or morpheme boundary, and it has proven useful in the recognition system.

For acoustic modeling we use the phone in context as basic unit. A word in the lexicon is then acoustically modeled by concatenating the phone models according to the phonemic transcription in the lexicon. The phone set definition for each language, as well as its consistent use for transcription is directly related to the acoustic modeling accuracy.

Gender-dependent models were estimated using the segmental MAP algorithm[6]. The first decoding pass with a bigram LM is used to generate a word lattice. The word lattice is generalized and reduced to form a word graph (without timing information) which is used to limit the search space for the trigram pass[2] with more accurate acoustic models. The second pass was run with all three model sets, retaining the hypothesis with the highest likelihood. For American English, British English, and German (German for EVAL-SYSTEM only) tied-state models were constructed in a manner similar to that used by [12]. The use of state-tying reduced the word error by 5-10% on the dry-run data which was subsequently used for system developement. For British English the acoustic model set contained 3x2582 tied states. The American English acoustic model set contained 3x2814 tied states. The German acoustic model set contained 3x3141 tied states (in the EVAL-SYSTEM, versus 883 CD models in the DRY-SYSTEM). The French acoustic model set contained 3x779 CD models as state-tying did not provide any significant performance improvement.

[2]The Phondat Corpus is available for research purposes from U. Munich.

EVALUATION

The SQALE test data consist of 200 sentences for each language (10 sentences from each of 20 speakers) for the dry-run (Feb95) and for the evaluation (May95) tests.[3] LIMSI's results for both test sets are given in Table 3. For the dry-run test data we also include results obtained with the final systems for each language in order to demonstrate the improvements obtained during system development. Concerning the dry-run test set we can note the high OOV rate (out of vocabulary words) in German compared to the other languages. German has 4.8% OOVs in the dry-run test for the 40k system while the other language OOV rates are below 2%. With the 64k German evaluation system the dry-run data has an OOV rate of 2.4%. The evaluation test set has a more balanced OOV rate across languages (ranging from 1.5 to 2.0%). The use of a higher n-gram LM results in a larger error reduction for the language with the best lexical coverage (English). Languages with a larger lexical variability require larger training text sets in order to achieve the same modeling accuracy. Improvements between the dry-run and evaluation systems can be measured comparing 'dry-tg' results in Table 3. Improvements of about 20% were obtained for the two new languages (British English and German). For British English these are mainly due to a new acoustic analysis, a larger number of context-dependent models using a tied-state approach for parameter sharing, and the use of gender-dependent models. In German additional error reduction comes from a sum of modifications, the relative contribution of each being difficult to estimate. Increasing the lexicon size from 40k to 64k, reduced the errors due to OOV words. A different subset of training sentences was selected (for the dry-run the short stories were not used, whereas isolated letter utterances were included, for the EVAL-SYSTEM we did the opposite), the phone set was reduced, glottal stop was optionally used for words starting with a vowel, phonetic transcriptions in the lexicon have been checked for consistency (this is particu-

[3]The French and British English test data were selected by TNO from the test portions of the BREF and WSJCAM0 corpora. The American English test data came from unused portions of the ARPA WSJ corpus, with additional sentences recorded by TNO. Since no read newspaper text corpus was available in German, TNO recorded all of the German test data[11].

Corpus	*WSJ*	*WSJCAM0*	*BREF*	*Phondat*
language	*American*	*British*	*French*	*German*
# training speakers	84	90	80	155
# training utterances	7k	7k	5.1k	16.5k
#distinct phones	46	45	35	51/48
#CD models	2390	2558	779	2481

Table 2: Comparison of speech corpora used for training, the number of phones and acoustic models in the EVAL-SYSTEM.

Corpus	*WSJ*	*WSJCAM0*	*BREF-LeMonde*	*Phondat-FR*
language	*American*	*British*	*French*	*German*
Lexicon size	20k	20k	20k	40k/64k
OOV-rate (dry)	1.2	1.3	1.9	4.8/2.4
OOV-rate (eval)	1.5	1.7	1.8	-/2.0
DRY-SYSTEM dry-bg	16.3	19.0	20.3	31.9
DRY-SYSTEM dry-tg	11.6	16.2	15.8	28.4
EVAL-SYSTEM dry-tg	11.5	13.1	14.7	21.8
EVAL-SYSTEM eval-bg	17.2	18.8	17.7	18.4
EVAL-SYSTEM eval-tg	13.5	15.4	15.3	16.1

Table 3: Recognition results are shown as %Werr (%Werr = %subs+%del+%ins) using DRY and EVAL systems with bigram and trigram LMs. OOV rates are given for the two test sets.

larly important in German due to the high compound-word rate), new transcriptions have been added. For all languages improvements come from enhanced acoustic modeling as the language models were unchanged.

SUMMARY

We have already demonstrated large vocabulary, continuous speech recognition systems for American English and French[3]. The American English system has been evaluated in the last 4 ARPA organized evaluations[2, 5] and a comparable system has been evaluated under closely matched conditions for French[3, 4]. In the context of the LRE SQALE project we have developed recognizers for British English and German. While the basic recognition technology is the same for all systems (phone-based CDHMM for acoustic modeling and n-gram statistics estimated on newspaper texts for language modeling), differences among the languages have been taken into account. For example, we have observed that the lexical coverage for German and French is less than that of English, due primarily to the large number of inflectional forms found in both these languages. In German compounding is an OOV source (like proper names for all languages). For this reason, the recognition vocabulary for German was larger than for the other languages. We have reported the SQALE dry-run and evaluation tests for all four languages. The evaluation system trigram results were 13.5% for American English, 15.4% for British English, 15.3% for French and 16.1% German, from which we can conclude that under somewhat comparable conditions (same number of test speakers, similar OOV rates, fixed acoustic and LM training data) the systems obtain word errors in the same range. We have previously demonstrated that lower word accuracies can be obtained by using more training data and/or larger recognition lexicon[3, 5]. For example, our 1994 *WSJ* 20k trigram system had a word error of 12.1%. The word error was reduced to 9.2% by use of a 65k trigram, which highlights the importance of lexical coverage in error reduction.

REFERENCES

[1] J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, " Speaker-Independent Continuous Speech Dictation," *Eurospeech-93*.

[2] J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker., "The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task," *ICASSP-94*.

[3] J.L. Gauvain,, L. Lamel, G. Adda, M. Adda-Decker, "Speaker-Independent Continuous Speech Dictation," *Speech Communication*, **15**, (1-2), 1994.

[4] J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, "Continuous Speech Dictation in French," *ICSLP-94*.

[5] J.L. Gauvain, L. Lamel, M. Adda-Decker, "Developments in Continuous Speech Dictation using the ARPA WSJ Task," *Proceedings ICASSP-95*.

[6] J.L. Gauvain, C.H. Lee, "MAP Estimation of Continuous Density HMM: Theory and Applications," *DARPA Sp. & Nat. Lang. Workshop*, Feb. 1992.

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

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

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

[10] T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals, "WSJ-CAM): A British English Speech COrpus for Large Vocabulary Continuous Speech Recognition," *ICASSP-94*.

[11] H.J.M. Steeneken, D.A. Van Leeuwen, "Multi-Lingual Assessment of Speaker Independent Large Vocabulary Speech-Recognition Systems: the SQALE Project, *Eurospeech'95*.

[12] S. Young, P. Woodland ,"State clustering in hidden Markov model-based continuous speech recognition," Computer Speech & Language **8**, 1994.