SPEECH RECOGNITION OF EUROPEAN LANGUAGES

Lori LAMEL

Spoken Language Processing Group
LIMSI-CNRS
BP 133
91403 Orsay France
email: lamel@limsi.fr

Renato DE MORI

School of Computer Science
McGill University
3480, University Street
Montreal, Quebec, Canada, H3A 2A7
e-mail: demori@cs.mcgill.ca

ABSTRACT

A basic overview is presented of the main ongoing efforts in large vocabulary, continuous speech recognition (LVCSR) for European languages. We address issues in acoustic modeling, lexical representation, and language modeling for several European languages, as well as issues in comparative evaluation.

1. INTRODUCTION

In this paper we aim to provide a basic overview of the main ongoing efforts in large vocabulary, continuous speech recognition for European languages. We address issues in acoustic modeling, lexical representation, and language modeling. In addition to presenting a snapshot of speech recognition in different European languages, we try to highlight language specific characteristics that must be taken into account in developing a recognition system for a given language. Some other issues that are touched upon are the availability of training and testing data in the different languages, the choice of recognition units, lexical coverage, language specificities such as the frequencies of homophones, monophone words, compounding, liaison, and phonological effects (reduction), and multilingual evaluation.

2. ACTIVITY OVERVIEW

Interesting results have been obtained in many languages using Hidden Markov Models (HMM) for phonetic units. The number and type of units modeled for achieving comparable performances on similar tasks, varies in different languages. A brief overview of active groups and models used in various Eurpean countries for large vocabulary ASR is presented below. The report is based on comments provided by colleagues who kindly accepted our request to provide information on activities in large vocaulary, continuous speech recognition in their countries. Activities in several languages, such as Dutch, Finnish, Danish, Slavic Languages and Greek are not reported because precise data on large vocabulary ASR were not available to us. We expect interesting results for these languages to be published in the near future.

Most of the ongoing work in English has made use of widely available corpora in American English, and work on this language us carried out by sites all over the world. The sites developing systems for the ARPA-sponosored *Wall Street Journal* and *North American Business News* evaluations include: AT&T, BBN, BU, CMU, CUED, CRIM, Dragon Systems, IBM, Karlsruhe University, LIMSI-CNRS, MIT Lincoln Labs, NYU, Philips and SRI.

The main sites working on large vocabulary, continuous speech recognition of French are in France (CNET, CRIN, IRIT, LIMSI-CNRS) and in Canada (CRIM), and IBM and BNR, (LV, but not CSR). In the context of the LRE SQALE project CUED and Philips have developed systems for French. The activities on the French language are likely to increase as a result of the recently created Francil Network, which will organize periodic evaluations on the French language.

Various groups are working on the German Language, among them, University of Erlangen (developed polyphone HMMs), Philips and the University of Aachen (interact in the development of large vocabulary dictation system), University of Karlsruhe (specialized in spoken language translation). In the context of the LRE SQALE project CUED and LIMSI have developed systems for German. A large activity in spoken language translation is under way in the Verbmobil project, sponsored by the German Ministry of Science and Technology (http://werner@ira.uka.de/ http://www.dfki.uni-sb.de/verbmobil). This project includes a large scale recognition and evaluation effort using large spontaneous speech corpora for translation of scheduling conversations. The effort involves other European partners in France, Italy and the UK in a Consortium for Speech Translation and Research (C-STAR).

The major groups active in Italy are CSELT and Politecnico di Torino, IRST and IBM Italy. The Italian counterpart of the Wall Street Journal, "il sole 24 ore" is used at IRST for LVCSR dictation experiments. An Italian version of ATIS is also available.

Activity in Portuguese is in progress at INESC and will be reported when results will be available.

There are few groups working in European Spanish Continuous Speech Recognition: Universidad Politecnica de Valencia, Universidad Politecnica de Madrid, Universidad Politecnica de Catalunya, and Universidad del Pais Vasco. IBM-Sevilla has produced the Spanish version of an isolated word Dictation Machine.

A major research effort in Sweden is located at the Royal Institute of Technology (KTH), where they have developed the Waxholm spoken language system. The Swedish PTT is also active in the speech technology field. The SRI-SICS-Telia Research Spoken Language Translator is a cooperative project to develop a speech-to-speech translator between Swedish and English in the ATIS domain.

3. ACOUSTIC MODELING

Acoustic modelling is performed with HMMs in almost all systems. Different languages have different sets of units and different coarticulation influences amomg adjacent phonemes. This influ-

ences the way of choosing context-dependent models and of tying distributions.

ENGLISH: American English is the language for which the largest amount of data for model training is available and in which extensive experiments have been conducted on various forms of parameter smoothing, context clustering and distribution tying. It appears that comparable systems work for British English as demonstrated in the SQALE project using the WSJCAM0 corpus. English is usually represented with a set of 40-50 phonemes, although in some systems finer phonetic distinctions are made.

FRENCH: For what concerns signal processing, it has been found that using a 4kHz bandwidth and an 8kHz bandwidth has no significant difference in recognition performance, either at the phone or word level. This is in contrast to English (and probably other languages) where the higher bandwidth typically brings a performance improvement to both the phone and word accuracies.

French has about 35 phonemes, with 14 vowels (3 nasal) and 20 consonants. Most LVCSR systems in French makes use of HMMs for acoustic modeling, with context-dependent (CD) phone models. For close to real-time systems, reduced CD model sets or context-independent (CI) models may be used. There is interest by some sites (CRIN) in multilevel systems where certain features may be used to hypothesize phone sequences, which are then passed to the LM component.

GERMAN: German has a large number of vowels (25) often including lax, tense normal and tense long distinctions for the same vowel type. Vowel¸-initial words and morphemes are often, but not systematically, preceded by a glottal stop or by glottalization. German also has a large number of consonant clusters that are subject to reduction at word boundaries. CD and CI phone modeling have been used in speech recognition.

ITALIAN: Italian has only seven vowels that can be reduced to five for practical purposes. The number of consonant clusters is not as large as in English. Detailed phoneme recognition experiments conducted so far at IRST with the Italian corpus APASCI result in 75% phoneme recognition with CI models and 82% with CD models. Experiments on continuous speech dictation on a 10000 word vocabulary task show a minor improvement on word recognition with CD models (91.17%) with respect to CI ones (88.47%).

SPANISH: Acoustic modeling of phone units is mostly based on CI Semi-Continuous and Discrete Hidden Markov Models. The latter are also used with Stochastic Regular Grammars. The best speaker-independent result achieved was 66% phone recognition with HMM+MLP and using only a bigram model of phones. Spanish is a phonetic language which has a reduced set of rules for a orthographic-phonetic transcription. Work is in progress to use context dependent models.

SWEDISH: The vowel system in Swedish is very rich. As many as 18 vowels can be identified. There are long/short consonants, some with considerable length duration (e.g. more than 100 ms occlusion). Endpoint detectors for English system tend to activate when a Swede tries to train or test a system with an English origin. The Swedish /r/ has many different allophones which cause problems. Retroflexation and nasality spread over syllable and word boundaries. Swedish is fairly clearly pronounced compared to English. Swedish is a tone language with two tones accent I and accent II. In a few cases this can cause confusions but, more important, the tone is a clear cue for the underlying stress pattern including signals for compunding. Prosody plays an important role in communication but has so far not been explored much in recognition.

4. LEXICAL REPRESENTATION

Lexicons are typically represented with phone-like symbols. The lexicon is the link between the acoustic models and the language model. In building the recognizer graph, each word is expanded according to its lexical pronunciation(s). Efficient tree lexical structures have been developed for the several languages[2, 9]. Words are often composed by more than one morpheme. This aspect has more complex realizations in certain languages. Differences in pronunciations are more frequent in certain languages than in others. Furthermore, they have different impacts on recognition accuracy.

FRENCH, ITALIAN, SPANISH: A common feature of these languages is that they have rich sets of terminations for verbs. A large proportion of verbs are regular, so there are repetitive structures that are common termination of many verbs.

GERMAN: Some peculiarities of German speech recognition systems are: (1) many inflexions of words, hence many different endings of words which require the use of a large lexicon and are hard to recognize; (2) a strong tendency to create compound words which enlarge the lexicon, uncompouding requires sophisticated morphological decomposition[4]; (3) long distance agreement a word like 'abfahren' (to depart) will be split in a sentence as 'ich fahre um 9 Uhr ab' (I will depart at 9 o'clock) which affects language modeling (4) a good correspondence between spelling and pronunciation (which may be used for word recognition training using only the orthographic transcription); (5) German has large dialectal variations, as described in a recent paper on automatic learning of these variations[8].

PORTUGUESE: European Portuguese is regarded as a very difficult language for foreign students to understand, due to the high degree of vocalic reduction. It is even difficult for Brasilians, where this type of pheneomena is not so pronounced. Activity in phoneme modeling is in progress.

SWEDISH: Swedish nouns have something corresponding to gender which influence suffixes and choice of articles. Swedish verb forms used to be quite a bit more complicated but have been simplified during the last 50 years[14]. has many compound words which are sometimes difficult to decompose. This creates problems both for speech understanding and speech synthesis.

4.1. Comparative Lexical Coverage

Various group are working on LVCSR using very large corpora for deriving the parameters of a statistical language models. Table 1 shows the characterisitcs of popular corpora in different languages.

Corpus language	WSJ English	Le Monde French	FR German	Sole 24 Italian
Training text size	37.2M	37.7M	25.7M	
#distinct words	165k	280k	650k	200k
5k coverage	90.6%	85.2%	82.9%	88.3%
20k coverage	97.5%	94.7%	90.0%	96.3%
40k coverage	99.2%	97.6%	-	98.8%
65k coverage	99.6%	98.3%	95.1%	99.0%
20k-OOV rate	2.5%	5.3%	10.0%	3.7%

Table 1: Comparison of *WSJ*, *Le Monde*, *Frankfurter Rundschau* and *Il Sole 24 Ore* lexica and LM training corpora.

4.2. Monophones

In French it is frequent to have words represented by a single phoneme. This is in contrast with English, where only vowels can be a monophone word (each of which has several orthographic forms). The number of monophones and the fact that they are frequent has several implications. First, rare words and OOVs are often easily replaced by a sequence of monophone words (all of which are pretty probable - with high backoff LM scores), so the recognized word string is phonemically correct even though it is orthographically wrong. The second problem is that these monophone words greatly increase the size of the recognition graph when crossword CD models are used. (For the LIMSI system the French graph takes about 1.5 to 2 times as much memory as for the same sized vocabulary system in English.)

4.3. Homophones

The problem of having words with different orthography but the same phonetic representation increases the complexity of the recognition task. Table 2 gives lexical homophone data for some popular speech corpora used in Europe. It appears as expected that the problem is less important in Italian than, for example, in French. In running text the homophone frequencies are even higher - 57% for French and 18% for English.

	Rate	Homophone class size (k)			
Corpus	*Lexicon*	*1*	*2*	*3*	*≥ 4*
BREF (10k)	35%	6686	1329	215	73
BREF (40k)	45%	23.7k	5361	1264	1039
WSJ (9k)	6%	8453	237	22	1
WSJ (65k)	15%	60.4k	3689	890	291
FR (64k)	10%	58.1k	2769	221	57
S24o (10k)	1.7%	9872	85	3	0

Table 2: Left: Single word homophones in French (BREF), English (WSJ), German (FR) and Italian (S24o). Right: Table entries correspond to the number of homophone classes with *k* graphemic forms in the class.

5. LANGUAGE MODELING

Most sites make use of statistical n-gram Language Models (LM), which are more or less efficient in the different languages. For languages in which agreement can span several words (like French, Italian, Spanish...), higher order n-grams than bigram and trigrams may be needed. This requires substantially more LM training text materials.

FRENCH: N-gram LMs have been successfully used for French. It has been demonstrated that a bigram LM is not strong enough to account for agreement. For example, the "ne" in "ne VERB pas" can be easily deleted with a bigram LM, but less so with a trigram LM.

GERMAN: Important results have been obtained in statistical language modelling with bigrams and polygrams[7],and phrase clustering[6].

ITALIAN: Various types of bigram mnd trigram models have been developed. Stochastic and non-stochastic context-free grammars have also been used for specific applications with medium and large size vocabularies[9]. Grammar constructs have also been used[10].

SPANISH: Spanish LMs have been developed using Stochastic Regular Grammars automatically learned by using Grammatical Inference Techniques. The task used was an oral query to a Spanish Geographical Data Base. The aplication of techniques based on N-grams to Spanish have also been explored for the same task.

SWEDISH: Swedish is a Germanic language and thus the grammar has similarities with both German and English. The Waxholm project uses the STINA parser to model both dialog and grammar[13]. The translation project is based on the SRI Core language Engine[15].

6. SPEECH RECOGNITION EVALUATION

The most widely known evaluation acitivites of large vocabulary, continuous speech recognition systems are those carried out under in the ARPA CSR program, starting with the Resource Management task (1000 words, word pair grammar), to the Wall Street Journal task (originally 5000 and 20,000 word vocabulary tests) to the North American Business News tests with unlimited vocabulary size. The commonly used measures of performance are the word error and the sentence error, as well as statistical measures to assess the significance in performance of the different systems. The word error is computed after performing dynamic programming alignment between the reference and hypothesized strings. In order to improve the alignment, NIST has proposed a phone-mediated alignment[17]. However, none of the measures take into account the type of error or the similarity of the words, or any particular analysis of errors due to out-of-vocabulary words.

In the LRE SQALE (Speech recognizer Quality Assessment for Linguistic Engineering) project this evaluation methodolgy was applied to multilingual evaluation, in order to experiment with establishing an evaluation paradigm in Europe for the assessment of large-vocabulary, continuous speech recognition systems. The recognition systems were tested using commonly agreed upon protocols (fixed training data, fixed vocabulary and language model), with the tests organized by the coordinating laboratory TNO. Multiple sites tested their algorithms on the same database, so as to compare the merits of different methods, and each system was evaluated on at least two languages, so as to compare the relative difficulties of the languages, and the degree of independency of the algorithm to a given language. The evaluation was carried out for American English (ARPA WSJ task), British English (WSJ-CAM0), French (Bref/*Le Monde*) and German (Phondat/*Frankfurter Rundschau*, using publicly available corpora. For English and German a 20k word vocabulary was used and for German a 64k word vocabulary was used in order to increase the lexical coverage. The test data was selected by TNO in order to have a comparable out-of-vocabulary word rate for all languages. The recognition performances obtained across languages and systems were somewhat comparable[20], which demonstrated that the same recognition technolgy developed for American English could be ported with reasonable success to other European languages.

7. DISCUSSION

It is still very difficult to make cross-language comparisons of speech recognition techniques, because there are some many uncontrolled differences. The use of large corpora is relatively recent and the corpora used (or available) in different languages have different characteristics. Furthermore, the speech data vary a lot in

quantity, variety, and the way they were collected. The interesting thing is that in spite of all these difficulties, performances for comparable task conditions are somewhat comparable, as demonstrated in the LRE SQALE evaluations.

Some observed commonalities across languages are that:

- more training data (acoustic or LM) improves performance;

- for acoustic training data from more speakers is generally better than fewer speakers, data with varied phone contexts is generally more easily ported to other tasks;

- gender dependent modeling usually gains something

- larger lexicons have lower OOVs and therefore higher recognition accuracy (this is the case for English, French and German 20k vs 40k or 60k)

- large variations in recognition accuracy observed across speakers (some of this variability can be reduced using quick adaptation techniques)

- while prosody and lexical stress are known to be important for some languages, these have not typically been exploited in current LVCSR systems.

Bearing in mind differences in corpora and test data, on the same test set we observe that phone recognition is better in French than English, but word recognition is better in English. It appears that the phonemes are produced more consistently in French, perhaps to counter-balance the inherent lexical ambiguity. In English, there are fewer homophones, and therefore speakers can be more variable in the realizations of the phones. We expect that phone recognition in Spanish and Italian should be easier than in Swedish and German.

Open issues for research in multilingual LVCSR include how to define well-balanced corpora, what are the optimal characteristics for new corpora, how to carry out multilingual evaluation, i.e., how can we compare the same, or harder yet, different recognition technology in different languages?

8. ACKNOWLEDGMENTS

9. REFERENCES

These references do not attempt to be complete, but suggestive of work being carried out in LVCSR in different European languages.
FRENCH:

[1] M. Dymetman, J. Brousseau, G. Foster, P. Isabelle, Y. Normandin, P. Plamondon "Towards an Automatic Dictation System for Translators: the Trans Talk Project", *ICSLP'94*.

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

[3] J.L. Gauvain, L.F. Lamel, G. Adda, and J. Mariani, "Speech-to-Text Conversion in French," *Int. J. Pattern Recognition & A.I.*, **8**, No. 1, 1994.

GERMAN:

[4] P. Geutner (1995), "Using morphology towards better large vocabulary speech recognition," *ICASSP-95*.

[5] R. Haeb-Umbach, H. Ney, "Improvements in time-synchronous beam-search for 10000-word continuous speech recognition," *IEEE Trans. SAP*, **2**, 353-365, April 1994.

[6] R. Kneser, H. Ney "Improved clustering techniques for class based statistical language modeling," *Eurospeech'93*.

[7] E.G. Schukat-Talamazzini T. Kuhn, H. Niemann, "Progress and Prospects of Speech Research and Technology," in H. Niemann, R. DeMori and G. Hahnrieder, Editors *Speech Recognition for Spoken Dialog Systems*, 1994.

[8] T. Sloboda, " Dictionary learning: Performance through consistence," *ICASSP-95*.

ITALIAN:

[9] M. Federico, M. Cettolo, F. Brugnara, G. Antoniol (1995), "Language Modeling for Efficient Beam-Search," *Computer Speech & Language* (in press)

[10] E. Giachin, "Phrase Bigrams for continuous speech recognition *ICASSP95* **1**.

SPANISH:

[11] G. Bordel, I. Torres, "QWI: A Method for Improved Smoothing in Language Modelling," *ICASSP-95*.

[12] E. Vidal, F. Casacuberta, P. Garcia (1995), "Syntactic Learning Techniques in Language Modeling and Acoustic-Phonetic Decoding", in "New Advances and Trends in Speech Recognition and Coding", NATO-ASI Series. Sprnger-Verlag. (In press).

SWEDISH: (internet consultation : http://www.speech.kth.se)

[13] M. Blomberg, R. Carlson, K. Elenius, B. Granström, J. Gustafson, S. Hunnicutt, R. Lindell, L. Neovius "An experimental dialogue system: WAXHOLM," *Eurospeech'93*, **3**.

[14] R. Carlson, K. Elenius, B. Granstrom, S. Hunnicutt (1986), "Phonetic properties of the basic vocabulary of five European languages: Implications for speech recognition," *ICASSP-86* **4**.

[15] M. Rayner, L. Carter, P. Price, B. Lyberg, "Estimating Performance of Pipelined Spoken Language Translation Systems," *ICSLP'94*.

MULTILINGUAL EVALUATION:

[16] C. Dugast, X. Aubert, R. Kneser, "The Philips Large-Vocabulary Recognition System for American English, French and German," *Eurospeech'95*.

[17] W. Fisher, J. Fiscus, A. Martin, D. Pallett, M. Przybocki, "Further Studies in Phonological Scoring" *ARPA Spoken Language Technology Workshop*, Austin, Tx, Jan. 1995.

[18] L. Lamel, M. Adda-Decker, J.L. Gauvain, "Issues in Large Vocabulary, Multilingual Speech Recognition," *Eurospeech'95*.

[19] D. Pye, S.J. Young, P. Woodland, "Large Vocabulary Multilingual Speech Recognition using HTK," *Eurospeech'95*.

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