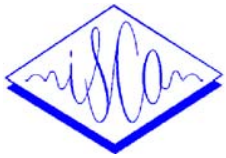

4th International Conference on Spoken Language Processing (ICSLP 96)
Philadelphia, PA, USA
October 3-6, 1996

Spoken Language Processing in a Multilingual Context

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

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

ABSTRACT

In this paper we overview the spoken language processing activities at LIMSI, which are carried out in a multilingual framework. These activities include speech-to-text conversion, spoken language systems for information retrieval, speaker and language recognition, and speech response. The Spoken Language Processing Group has also been actively involved in corpora development and evaluation. The group has regularly participated in evaluations organized by ARPA, in the LE-SQALE project, and in the AUPELF-UREF program for provision of linguistic resources and evaluation tests for French.

1. INTRODUCTION

A large number of research activities are presently being explored at LIMSI concerning spoken language processing in a multilingual context. These activities include multilingual, large vocabulary, speaker-independent continuous speech dictation [5, 4, 2, 3], the development of multilingual spoken language systems [19, 10, 8], automatic speaker and language identification [20, 13, 21]. Investigations in automatic prosodic feature extraction [14] and stochastic concept modeling [11] aim at different levels of representation, to improve spontaneous speech recognition and understanding.

Our present read speech recognition systems achieve comparable recognition accuracies in different European languages (French, German and British English) and in American English [4]. They are speaker-independent and allow for adaptation to improve individual speaker's recognition scores. Robustness of the developed systems to environmental noise and telephone channel is an important issue being investigated [2].

A challenging research domain deals with the adaptation of state-of-the-art laboratory read speech recognizers for use in real applications. Real applications may impose requirements on the system such as limitations on available computational resources, the need for real-time processing, and acoustic modeling to deal with background noise. Many potential applications entail spontaneous speech processing, spoken language understanding, response generation and dialog management. Speech-to-text research also serves as the technology foundation for other related activities, such as speaker and language identification.

Much of the demonstrated progress in speech recognition and spoken language understanding over recent years has been fostered by the availability of large commonly used corpora for system training and evaluation in different languages. Corpora development is one of our important supporting activities. We record subjects on a regular basis for a variety of tasks (read newspaper text, telephone-based collection for multiple languages, spoken language information retrieval using a laboratory kiosk and by telephone).

2. MULTILINGUAL SPEECH DICTATION

The applicability of speech recognition techniques for different languages is of particular importance in Europe, where each country has its own national language. During the last years we have extended our speech recognition system to different languages, including now American English, French, British English and German. Most of our large vocabulary, continuous speech recognition research (LVCSR) in English focuses on the ARPA Wall Street Journal (WSJ) and North American Business News (NAB) tasks[16, 2]. LIMSI has participated in annual ARPA continuous speech recognition benchmark tests since 1992, consistently demonstrating state-of-the-art recognition performance. For French this research relies heavily on the BREF speech corpus[17] for acoustic model training and 50 million words of text from the French newspaper *Le Monde* for the language model training material[16]. In the context of AUPELF-UREF we are participating in the elaboration of an evaluation protocol and infrastructure for evaluation of speech technologies developed for the French language. Within the LE-SQALE project our recognizer has been adapted to British English using the WSJCAM0 corpus [24], and to German using PHONDAT [1] corpus [22].

As the recognizer makes use of phone-based continuous density HMM for acoustic modeling and n-gram statistics estimated on newspaper texts for language modeling, its adaptation to a new language consists mainly in the creation of the language specific acoustic and language models. This can appear as a rather straightforward process, once you have at your disposal sufficient speech and text databases. Obtaining appropriate text and speech corpora is thus an important issue to train reliable acoustic and language models for a new language.

2.1. Text and Speech Corpora

The importance of large corpora for training acoustic and language models cannot be overemphasized. Much of the recent progress is largely due to the availability of large corpora for training and testing speech recognition technology in a multilingual context.

For dictation tasks, it is relatively easy to obtain text data for training language models for most languages under consideration. Most of the automatic processing of the texts for further use appears to be relatively language independent. To obtain appropriate speech data, a subset of texts can be selected to ensure good phonetic

[1] For LM, training texts from the *Frankfurter Rundschau* newspaper were obtained from the ACL-ECI CDROM.

10.21437/ICSLP.1996-558

coverage and used as prompts to collect spoken data for acoustic model training.

Concerning French LIMSI has been actively involved in data collection, with some of the efforts partially sponsored by national or European projects. We have designed and recorded the BREF-*Le Monde* read newspaper speech corpus[17]. In the context of the AUPELF-UREF evaluations, we are processing additional *Le Monde* texts and will record new development and test utterances.

Table 1 compares different newspaper text corpora from different countries in terms of lexical variety and lexical coverage achieved for different sized lexica.

Corpus	*WSJ*	*Le Monde*	*FR*	*Sole 24*
language	*English*	*French*	*German*	*Italian*
Training text size	37.2M	37.7M	36M	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%
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* text corpora in terms of number of distinct words and lexical coverage for different lexicon sizes.

There is a certain amount of text normalisation to be carried out, in order to clean the texts and to define what is actually to be considered a lexical item in each language. Once normalised a task vocabulary can be selected and language models trained. A common normalisation motivation for all languages is the reduction of lexical variability in order to increase coverage for a fixed size task vocabulary. Reducing lexical variety may give higher lexical coverage and more robust language model, but with a loss in syntactic or semantic resolution. Normalisation decisions can be language-specific. In English it is common to capitalize all the texts, and thus no lexical distinction is made between (*Gates*, *gates*), (*Green*, *green*) for instance. In French capital letters are kept distinctive for proper names, resulting in different lexical entries for (*Pierre*, *pierre*) or (*Roman*, *roman*) for example. In German all substantives are written with capitalized first letters and most words can be substantivized, thus generating lexical variability and homophones. But this kind of variability remains small. Measuring lexical coverage of case-sensitive and case-insensitive German texts yields a relative out-of-vocabulary (OOV) rate reduction of about 5% for 40k (from 6.8% to 6.5%) and 64k (from 4.9% to 4.7%) lexica respectively.

2.2. Lexical Representation

Elaborating, updating and improving a pronounciation lexicon for a new language is less straightforward and is based on language-specific knowledge, ranging from the definition of a language-specific phone set to the development of a grapheme-to-phoneme conversion system. The LIMSI lexicons are represented phonemically, using language-specific sets of phonemes. Alternate pronunciations are provided for about 10% of the words in French[2] and English. The pronunciation lexicons for French and American English were developed at LIMSI. For British English we combined

[2]This does not count word final optional phonemes marking possible liaisons for French. Including these raises their number to almost 40%.

```
French
les             le(C.) lez(V)
mon             mO mOn(V)
contenu         kOt{x}ny
autres          ot(C.) otrx otr(V) otrxz(V)
German
Instrument      ?In[sS]trumEnt
funktional      fUG{k}ts1onal
Zoologe         tso{?}ol0gX
```

Figure 1: Example lexical entries for French and German. Phones in {} are optional, phones in [ ] are alternates. () specify a context constraint, where **V** stands for vowel, **C** for consonant and the period represents silence. ? stands for glottal stop.

Corpus	*WSJ*	*WSJCAM0*	*BREF*	*Phondat*
language	*Am.*	*Br.*	*Fr.*	*Ger.*
#train. speakers	84	90	80	155
#train. utterances	7k	7k	5.1k	16.5k
#distinct phones	46	45	35	48
#CD models	2390	2558	779	2481

Table 2: Comparison of speech corpora used for training with the number of phones and acoustic models in the final system settings.

portions of the BEEP dictionary from Cambridge University with pronunciations taken from LIMSI American English WSJ lexicons, which were remapped to be more "British". For German, a 64k pronounciation lexicon was distributed by Philips within the SQALE project. The original phone set has been reduced and phonological variants have been added mainly to account for typical consonant cluster reduction or for some dialectal variation. About 5 % of the lexical entries in German allow multiple phonemic transcriptions. Some example entries for French and German are shown in Figure 1.

2.3. Multilingual Dictation Experiments

Our recognizer has been assessed on a newspaper dictation task in the context of the LE-SQALE project whose objective was to experiment with installing in Europe a multilingual evaluation paradigm for the assessment of large vocabulary, continuous speech recognizers[4]. A common vocabulary list and LM were 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 *Le Monde* training text corpus. Due to the significantly lower lexical coverage of German, a 64k-word vocabulary was chosen for the *Frankfurter Rundschau*. The common acoustic training data and LIMSI system characteristics are summarized in Table 2.

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] For both test sets, results with the final systems for each language are given in Table 3. Concerning the dry-run test data, there is a relatively high OOV rate in German with a 64k lexicon compared to the other languages with only a

[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[25].

Language	*Am.*	*Br.*	*Fr.*	*Ger.*
Lexicon size	20k	20k	20k	64k
%OOV (dry)	1.2	1.3	1.9	2.4
tg. (dry)	11.5	13.1	14.7	21.8
%OOV (eval)	1.5	1.7	1.8	2.0
trigram (eval)	13.5	15.4	15.3	16.1
bigram (eval)	17.2	18.8	17.7	18.4

Table 3: Recognition results are shown as %Werr (%Werr = %Subs+%Del+%Ins) using bigram and trigram LMs. OOV rates are given for the two test sets (dry & eval).

20k lexicon. The evaluation test set has a more balanced OOV rate across languages (ranging from 1.5 to 2.0%). Comparing bigram and trigram results shown for the eval set, we note that the use of a higher *n*-gram LM results in a larger error reduction for English, which has the best lexical coverage. Languages with a larger lexical variability require larger training text sets in order to achieve the same modeling accuracy, and may benefit more from longer span *n*-grams if there is sufficient training data. Another potential way to reduce lexical variability is by using more powerful, language-specific text normalizations. Doing so can result in lower OOV rates and more robust language models. In German, a major obstacle to high lexical coverage arises from word compounding for which morphological decomposition seems hold promise.

Better recognition results have been reported by training more accurate models by using additional acoustic and text data [1, 15, 23].

3. TOWARDS MULTILINGUAL SPONTANEOUS SPEECH RECOGNITION

When going from read to spontaneous speech, variability is added to the acoustic signal, in terms of speaking rate and style, speech disfluencies, new lexical items and syntactic structures. This increase in variability can be considered independent of the language under consideration, whereas the different forms of variability may be language-dependent.

Reasons of differences between read and spontaneous speech stem certainly in the difference in nature of both types of speech. Instead of an a priori written text, in spontaneous speech the speaker's intention has to be put in a linguistic oral form which is meant to be understood by a listener. The fact that the message is often composed in parallel with the speaking process, implies larger variations in speaking style, prosodics, and speaking rate, as illustrated by the French example in Figure 2. In this example, taken from the MASK corpus[7], the speaker asks about train information: *where does one change, ... where is the change*. The first part is uttered very rapidly with nearly whispered speech. The second part, spoken clearly at a normal speaking rate, repeats the same information using a different word sequence. Speech disfluencies (hesitations, uncomplete words..) and rearranging of word sequences or incorrect syntactic structures are often observed. In the example shown in Figure 3, a German speaker describes the house where he lives: ... *it is a family - , a 'more than one family' house....* The 3 word sequence *es ist ein* is reduced to a single syllabic cluster *sch'n*. The compound word *Familienhaus* is incompletely uttered as *Familie-* and then more precisely specified as *Mehrfamilienhaus*.

3.1. Speech and Text Corpora Development

In order to study characteristics of spontaneous speech and to have them accurately modeled, we are working on spontaneous speech corpus definition and collection [6, 19]. Data collection for spoken language systems generally concerns application-specific spontaneous speech, and often the vocabulary size can be relatively limited (on the order of 2000 words). As for conversational speech, acquiring sufficient amounts of language model training data is more challenging than obtaining acoustic data. With 10k queries relatively robust acoustic models can be trained, but these queries contain only on the order of 100k words which is insufficient for training *n*-gram language models. For spoken language systems, the most effective manner of obtaining representative data is collecting speech from users interacting with preliminary versions of a complete system. We have observed that as the system improves, subjects speak more easily and use longer and more varied sentences[7]. This also leads to the occurrence of more new words in the queries.

Spoken language corpora have been recorded for three information retrieval tasks in the travel domain: L'ATIS a French version of the ARPA ATIS task [8] (for information about fares and flights) the ESPRIT MASK (Multimodal-Multimedia Automated Service Kiosk) task [10] for access to rail travel information among 600 cities in France; and the LE MLAP RAILTEL task [9] to provide train timetable and scheduling information over the telephone. For L'ATIS we have recorded more than 10k queries from about 300 speakers, totaling about 110k words, less than 2k of which are distinct. MASK data collection has been carried out with three different systems, totaling at present about 30k queries with 200k words. About 100 of the speakers were recorded with the data collection system at the St. Lazare train station in Paris. We are now recording subjects on a data collection kiosk to better simulate the future MASK kiosk. Over 1000 calls have been recorded with the RAILTEL data collection system, for about 8k queries and 60k words. This summer we will start recording data for more general tourist information in the Paris area for use in the AUPELF spoken dialog action. Whereas the data collection is purely language-specific, the semantic frame representation of a given application, can be adapted in a multilingual context [12].

We have participated in the design and recordings of multilingual telephone based corpora in the context of national and European projects. For research in automatic language identification we have recorded a large, multilingual (French, English, German and Spanish) corpus of telephone speech. This corpus contains speech from 250 speakers of each language calling the LIMSI data collection system from their home country. We have observed different reactions of callers, for example, Spanish speakers are much more expressive in response to questions than their British colleagues.

Other corpora related activities include participation in the LRE

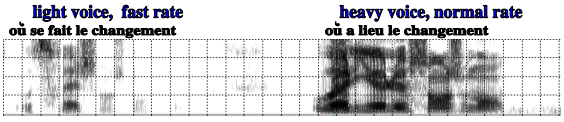

Figure 2: Spectrogram illustrating the spoken message elaboration process in spontaneous speech, resulting here in different speaking styles and information repetition. The French utterance is: *où se fait le changement, où a lieu le changement* (scale: 100ms x 1kHz, y-axis on 4kHz).

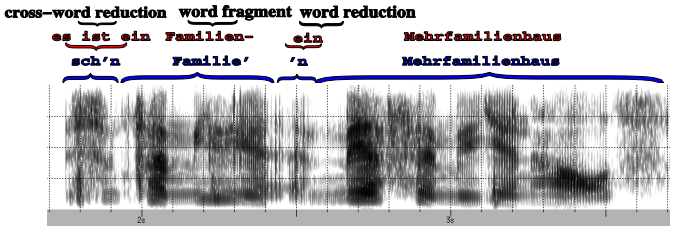

Figure 3: Spectrogram illustrating different typical problems of spontaneous speech. Word reduction, cross-word reduction, and word fragment phenomena can be observed. The language is German. (scale: 100ms x 1kHz, y-axis on 4kHz).

project RELATOR and the ESPRIT Network of Excellence ELSNET, preparation and contribution to the launching of the European Linguistic Resource Agency (ELRA), and in the Copernicus BABEL project aimed at creating comparable corpora in 5 Eastern European languages.

3.2. Spoken Language Systems

Our spoken language system (SLS)[19, 10] consists of a speaker-independent continuous speech recognizer, whose output is passed to a natural language (NL) component. The NL component is concerned with understanding the meaning of the spoken query and includes the semantic analysis [8] and dialog management. Natural language responses are automatically generated from the semantic frame, the dialog history and retrieved DBMS information, and synthesized using concatenated speech from stored dictionary units[18]. Information return can be only vocal (as by telephone) or both visual and vocal. Acoustic and language models of the recognizer are periodically reestimated using the additional collected training data. The semantic analysis is refined in light of the new corpora and system's understanding failures. The portability of the case frame approach between French and English has been shown [12]. Evaluating progressively our MASK SLS system on both word recognition level and SLS response level has proven to provide significant error reduction due to the additional training data supply. The word error rate went down from an initial rate of 22% to 10%.

4. CONCLUSIONS

Our experience with speech recognition of read speech has shown that the same recognizer can be adapted to different languages, provided that sufficient text and speech material are available for training the new language. The phone-based acoustic modeling approach requires a phonemic transcription lexicon, the elaboration and maintainance of which requires human expertise. Language dependencies were observed, for example, in French a large number of homophone and monophone words are responsible for a significant amount of decoding errors, whereas in German poorer recognition accuracy is due to lower lexical coverage (and as a result less precise language modeling) mainly stemming from word compounding.

Spontaneous speech processing is less portable to new languages as appropriate training data do not usually exist. While this is true at the acoustic level, text data for language modeling is even harder to obtain. Despite our experience as a community, constructing corpora that are representative, complete, and yet at the same time not too big, is an open research area. It is extremely hard to even demonstrate the effects of different corpus design strategies. Identical strategies applied in a multilingual context will not yield identical results.

In summary, our experience is that although general technologies and development strategies appear to port from one language to another, to obtain optimal performance language specificities must be taken into account.

REFERENCES

[1] J.L. Gauvain et al., "Speaker-Independent Continuous Speech Dictation," Speech Communication, **15**, Sept. 1994.

[2] J.L. Gauvain et al., "Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task," *ICASSP-96*.

[3] M. Adda-Decker et al. "Developments in Large Vocabulary, Continuous Speech Recognition of German," *ICASSP-96*.

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

[5] J.L. Gauvain et al. "Continuous Speech Dictation in French," *ICSLP'94*.

[6] A. Life et al., "Data Collection for the MASK Kiosk: WOz vs Prototype System," *ICSLP'96*.

[7] L. Lamel et al., "Development of Spoken Language Corpora for Travel Information," *Eurospeech'95*.

[8] S. Bennacef et.al, "A Spoken Language System For Information Retrieval," *ICSLP'94*.

[9] S.K. Bennacef et al., "Dialog in the RAILTEL Telephone-Based System," *ICSLP'96*.

[10] J.L. Gauvain et al., "The Spoken Language Component of the Mask Kiosk," *Human Comfort & Security Workshop*, Brussels, Oct. 26, 1995.

[11] W. Minker, S. Bennacef, J.L. Gauvain, "A Stochastic Case Frame Approach for Natural Language Understanding," *ICSLP'96*.

[12] W. Minker, S.K. Bennacef, "Compréhension et Évaluation dans le Domaine ATIS," *Journées d'Études en Parole, JEP'96*.

[13] L. Lamel, J.L. Gauvain, "A Phone-based Approach to Non-Linguistic Speech Feature Identification," *Computer Speech & Language*, **9**, pp. 87-103, Jan. 1995.

[14] E. Geoffrois, "Extraction robuste de paramètres prosodiques pour la reconnaissance de la parole", PhD thesis, U. Paris XI, 1995.

[15] J.L. Gauvain et al., "The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task," *ICASSP-94*.

[16] J.L. Gauvain et al., "The LIMSI Continuous Speech Dictation System," *ARPA HLT Workshop* 1994.

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

[18] L. Lamel et,al., "Generation and Synthesis of Broadcast Messages," *ESCA Workshop Applications of Speech Technology*, Lautrach, Germany, Sept. 1993.

[19] L. Lamel, S. Bennacef, H. Bonneau-Maynard, S. Rosset, J.L. Gauvain, "Recent Developments in Spoken Language Sytems for Information Retrieval," *ESCA Workshop on Spoken Dialog Systems*, Vigsø, Denmark, Spring 1995.

[20] L. Lamel, J.L. Gauvain, "Language Identification Using Phone-based Acoustic Likelihoods," *ICASSP-94*.

[21] J.L. Gauvain, L.F. Lamel, B. Prouts, "Experiments with Speaker Verification over the Telephone," *Eurospeech-95*.

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

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

[24] T. Robinson et,al. "WSJCAM): A British English Speech COrpus for Large Vocabulary Continuous Speech Recognition," *ICASSP-94*.

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