An Overview of Speech Recognition Activities at LIMSI

Jean-Luc Gauvain, Gilles Adda, Martine Adda-Decker, Claude Barras,
Langzhou Chen, Michèle Jardino, Lori Lamel, Holger Schwenk

Spoken Language Processing Group (http://www.limsi.fr/tlp)
LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France

ABSTRACT

This paper provides an overview of recent activities at LIMSI in multilingual speech recognition and its applications. The main goal of speech recognition is to provide a transcription of the speech signal as a sequence of words. Speech recognition is a core technology for most applications involving voice technology. The two main classes of applications currently addressed are transcription and indexation of broadcast data and spoken language dialog systems for information access.

Speaker-independent, large vocabulary, continuous speech recognition systems for different European languages (French, German and British English) and for American English and Mandarin Chinese have been developed. These systems rely on supporting research in acoustic-phonetic modeling, lexical modeling and language modeling.

1. INTRODUCTION

Speech recognition and related application areas have been a long term research topic at LIMSI, going back to the early 1980's. Our aim is to develop basic speech recognizer technology that is speaker-independent and task-independent, which at least means that using very large (ideally unlimited) recognition vocabularies. The systems also have to be robust with respect to background and channel noise and changes in microphone and microphone placement. They must have the ability to deal with spontaneous speech, including well-known disfluencies.

Speech recognition performance is acknowledged to be heavily dependent on the correctness of the acoustic and language models used, and the recognition lexicon. These are areas of active research in the group. The different sources of variability present in the speech signal is addressed by the use of different modeling techniques. However, what is considered non-pertinent for word recognition can be quite relevant from another perspective. A variable number of labels can be associated with the acoustic models (phoneme, talker's gender, identity, dialect, language, ...) and depending upon the application, the recognition system can eventually identify the acoustic or channel conditions, the speaker, and the language along with the linguistic content encoded in the speech signal.

In addition to our main research activites, substantial effort is devoted to system evaluation and to the development of speech corpora. Concerning evaluation, LIMSI was one of the first non-American labs to participate in annual DARPA evaluations of speech recognition technology and has regularly participated in benchmark tests since 1992, on tasks ranging from Resource Management to Wall Street journal and most recently, broadcast news transcription. These evaluations permit the comparison of different systems on the same test data, using common training materials and test protocols. The LIMSI system consistently obtained top level performance (always among the top 3 sites), achieving the highest word accuracy on four of the baseline tests. Evaluation of multilingual systems was carried out in the context of the LRE project SQALE, and the LIMSI recognition system achieved the lowest word error rate in the 1997 Aupelf sponsored evaluation of read-speech in French. In 1998 and 1999 we participated in the 1999 TREC-8 and 2000 TREC-9 SDR evaluation for retrieval of audio documents.

Research contracts cover most of the groups activities, the most recent European projects being: ALERT, CORETEX, ECHO, OLIVE, as well as French National projects from the DGA (Délégation Générale de l'Armement), and RNRT (Vocadis and Theoreme). The group also participates or has participated in several projects related to the development and distribution of linguistic resources and evaluation.

In the following sections we give a brief overview of our research activities related to developing core speech recognition technology and applying this technolgoy to various languages, and mention some of our recent research projects. Recent advances have been in techniques for robust acoustic feature extraction and normalization; improved training techniques which can take advantage of very large audio and textual corpora; algorithms for audio segmentation; unsupervised acoustic model adaptation; efficient decoding with long span language models; ability to use very large vocabularies. Much of recent progress can be linked to the availability of large speech and text corpora and simultaneous advances made in computational means and storage, which have facilitated the implementation of more complex models and algorithms.

2. CORE TECHNOLOGY

Speech recognition is principally concerned with the problem of transcribing the speech signal as a sequence of words. The LIMSI system, in common with most of today's state-of-the-art systems, makes use of statistical models of speech generation. From this point of view, message generation is represented by a language model which provides an estimate of the probability of any given word string, and the encoding of the message in the acoustic signal is represented by a probability density function (HMM). The speech decoding problem then consists of maximizing the *a posteriori* probability of the word string given the observed acoustic signal.

One of our primary objectives is to improve core technology for speaker-independent, continuous speech recognition so as to minimize the expected performance degradation under mismatched training/testing conditions. Our research thus focuses on extending the capabilities of the system to deal with unlimited-vocabulary speech in a variety of acoustical conditions, with different background environmental noise, unknown channels and microphones.

The LIMSI speaker-independent, large vocabulary, continuous speech recognizer makes use of continuous density hidden Markov models (HMMs) with Gaussian mixtures for acoustic modeling. The acoustic and language models are trained on large, representative corpora for each task and language [8, 12]. Each word is represented by one or more sequences of context-dependent phone models (intra and interword) as determined by its lexical transcription. The acoustic parameterization is based on a cepstral representation of the speech signal.

Word recognition is carried out in one or more decoding passes with more accurate acoustic and language models used in successive passes. For many applications there are limitations on the response time and the available computational resources, which in turn can significantly affect the design of the acoustic and language models. For each operating point, the right balance between model complexity and search speed must be found to optimize performance. A 4-gram single pass decoder which uses cross-word models has been implemented. The decoder makes use of a variety of techniques to reduce the search space and computation time, such as LM state conditioned lexicon trees, acoustic and language model lookahead, predictive pruning and fast Gaussian likelihood computation [7].

We have recently carried out some experiments to investigate the use of voting schemes to combine transcriptions produced by different speech recognizers. An extended ROVER algorithm has been developed that incorporates language model information and is of particular interest when the outputs from only a few recognizers are combined [26].

Speaker-independence is achieved by estimating the parameters of the acoustic models on large speech corpora containing data from a large speaker population (several hundreds to thousands of speakers). Local contextual variation is modeled by using large sets of context-dependent phone models, and the variability associated with the acoustic environment, the microphone and transmission channel are accounted for by adapting the acoustic models to the particular conditions or by using an explicit model of the channel.

Regularities and local syntactic constraints are captured via n-gram models, which attempt to account for the syntactic and semantic constraints by estimating the probability of a word in a sentence given the preceding *n*-1 words. Given a large corpus of texts (or transcriptions) it may seem relatively straightforward to construct *n*-gram language models. The main considerations are the choice of the vocabulary and the definition of words, such as the treatment of compound words or acronyms, and the choice of the backoff strategy. There is, however, a significant amount of effort needed to process the texts before they can be used. One motivation for normalization is to reduce lexical variability so as to increase the coverage for a fixed size task vocabulary. Normalization decisions are generally language-specific. For example, some standard processing steps include the expansion of numerical expressions, treatment of isolated letters and letter sequences, and optionally elimination of case distinction. Further semi-automatic processing is necessary to correct frequent errors inherent in the texts, and the expansion of abbreviations and acronyms. The error correction consists primarily of correcting obvious misspellings. Better language models are obtained by using texts transformed to be closer to the observed reading style, where the transformation rules and corresponding probabilities are automatically derived by aligning prompt texts with the transcriptions of the acoustic data [13].

An essential component of the transcription system is the recognition lexicon which provides the link between the lexical entries (usually words) used by the language model and the acoustic models. Lexical modeling consists of defining the recognition vocabulary and associating one or more phonetic transcriptions for each word in the vocabulary. Each lexical entry is described as a sequence of elementary units, usually phonemes. We explicitly represent silence, breath noise, and a filler words with specific symbols. The American English pronunciations are based on a 48 phone set, whereas for French and German sets of 37 and 49 phonemes are used, respectively. A pronunciation graph is associated with each word so as to allow for alternate pronunciations. For the Mandarin language a set of 40 phones is used. Since Mandarin is a tone-based language, with five different tones associated with the syllables, we investigate explicitly modeling tone in the pronunciation lexicon. For pronunciations with tone, we chose to distinguish only 3 tones: flat (tones 1 and 5), rising (tones 2 and 3), falling (tone 4).

3. PROCESSING AUDIO STREAMS

A major advance in speech recognition technology is the ability of todays systems to deal with non-homogeneous data as is exemplified by broadcast data. With the rapid expansion of different media sources for information dissemination, there is a pressing need for automatic processing of the audio data stream. A variety of near-term applications are possible such as audio data mining, selective dissemination of information, media monitoring services, disclosure of the information content and content-based indexation for digital libraries, etc. Broadcast news shows are challenging to transcribe as they contain signal segments of various acoustic and linguistic nature, with abrupt or gradual transitions between segments. The signal may be of studio quality or have been transmitted over a telephone or other noisy channel (ie., corrupted by additive noise and nonlinear distorsions), as well as speech over music and pure music segments. The speech is produced by a wide variety of speakers: news anchors and talk show hosts, reporters in remote locations, interviews with politicians and common people, unknown speakers, speakers with strong regional accents, non-native speakers, etc. The linguistic style ranges from prepared speech to spontaneous speech.

Two principle types of problems are encountered in transcribing broadcast news data: those relating to the varied acoustic properties of the signal, and those related to the linguistic properties of the speech. Problems associated with the acoustic signal properties are handled using appropriate signal analyses, by classifying the signal according to segment type and by training specific acoustic models for the different acoustic conditions.

Transcribing such data requires significantly higher processing power than what is needed to transcribe read speech data in a controlled environment, such as for speaker adapted dictation. Although it is usually assumed that processing time is not a major issue since computer power has been increasing continuously, it is also known that the amount of data appearing on information channels is increasing at a close rate. Therefore processing time is an important factor in making a speech transcription system viable for audio data mining and other related applications.

The LIMSI broadcast news automatic indexation system [11] consists of an audio partitioner [9], a speech recognizer [10, 12] and an indexation module [6]. The transcription components are shown in Figure 1.

Partitioning the Audio stream

The goal of audio partitioning is to divide the acoustic signal into homogeneous segments, removing non-speech segments, and labeling and structuring the acoustic content of the data. The result of the partitioning process is a set of speech segments with cluster, gender and telephone/wideband labels, which can be used to generate metadata annotations. While it is possible to transcribe the continuous stream of audio data without any prior segmentation, partitioning offers several advantages over this straight-forward solution. First, in addition to the transcription of what was said, other interesting information can be extracted such as the division into speaker turns and the speaker identities, and background acoustic conditions. Second, by clustering segments from the same speaker, acoustic model adaptation can be carried out on a per cluster basis, as opposed to on a single segment basis, thus providing more adaptation data. Third, prior segmentation can avoid problems caused by linguistic discontinuity at speaker changes. Fourth, by using acoustic models trained on particular acoustic conditions (such as wideband or telephone band), overall performance can be significantly improved. Finally, eliminating non-speech segments and dividing the data into shorter segments (which can still be several minutes long), substantially reduces the computation time and simplifies decoding.

The LIMSI partitioning approach relies on an audio stream mixture model [9]. Each component audio source, representing a speaker in a particular background and channel condition, is in turn modeled by a GMM. The segment boundaries and labels are jointly identified by an iterative maximum likelihood segmentation/clustering procedure using GMMs and agglomerative clustering.

Speaker and Language Recognition

Speaker and language recognition are a logical extension of continuous speech recognition as basically the same speech models can be used. The basic idea is to construct feature-specific model sets for each non-linguistic speech feature to be identified, and to process the unknown speech by all model sets in parallel. Instead of retaining the recognized string (as is done in recognition), what is of interest is which of the model sets has the highest likelihood. The feature associated with that model set is then attributed to the signal [17]. This is usually using phonotactic models, i.e. without the use of lexical information.

Another such feature which is commonly identified is the sex of the speaker, which is used to select sex-dependent models prior to word recognition. For high quality speech, identification of the speaker's sex is essentially perfect, and in the few cases where errors are made better recognition results are usually obtained with the chosen models (of the opposite sex).

Some references to our work on speaker and language identification can be found in [3, 18, 24].

Word transcription

For each speech segment, the word recognizer determines the sequence of words in the segment, associating start and end times and an optional confidence measure with each word. Word recognition is usually performed in two or more steps to allow unsupervised acoustic model adaptation. For audio stream data, the initial hypotheses are used in cluster-based acoustic model adaptation using

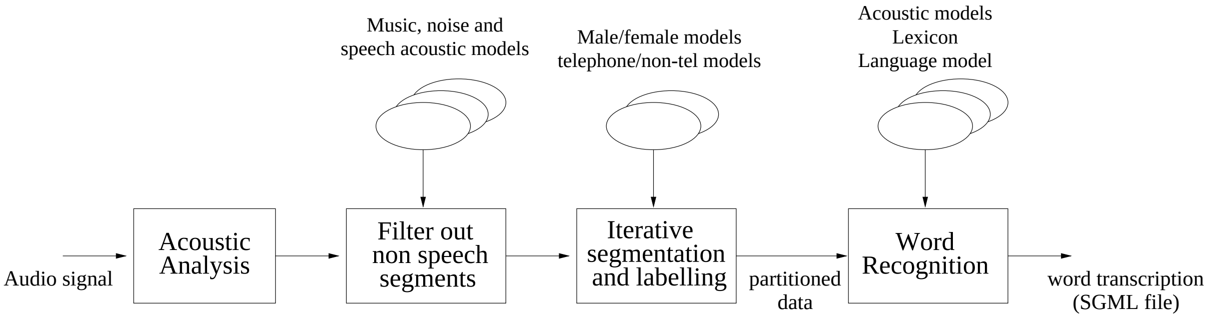

Figure 1: Overview of transcription system for audio stream.

the MLLR technique [22] prior to word graph generation, and all subsequent decoding passes.

Due to the availability of large corpora of audio and textual data for broadcast news in American English, most efforts in transcription have focused on this language. To evaluate the generalizability of the underlying methods, we have applied the partitioning and transcription algorithms developed for American English broadcast news data to three other languages. The development of the French and German systems has been partially financed by the the EC LE4-OLIVE project and by the French Ministry of Defense. The Mandarin language was chosen because it is quite different from the other languages (tone and syllable-based), and Mandarin resources are available via the LDC as well as reference performance results.

Current state-of-the-art laboratory systems can transcribe unrestricted American English broadcast news data with word error rates under 20%. Our transcription systems for French and German have comparable error rates for news broadcasts [1]. The character error rate for Mandarin is also about 20% [2]. Based on our experience, it appears that with appropriately trained models, recognizer performance is more dependent upon the type and source of data, than on the language. For example, we found that foreign documentaries are particularly challenging to transcribe, as the audio quality is often not too high, and there is a large proportion of voice over.

Spoken Document Retrieval

The automatically generated partition and word transcription can be used to for indexation and information retrieval purposes. Spoken document retrieval (SDR) can support random access to relevant portions of audio or video documents, reducing the time needed to identify recordings in large multimedia databases. For such applications, processing time is an important factor, and imposes constraints on the development of acoustic and language models. The LIMSI spoken document indexing and retrieval system combines a state-of-the-art speech recognizer [12] with a text-based IR system [6]. The same techniques commonly applied to automatic text indexation have been applied to automatic transcriptions of broadcast news radio and TV documents. Such techniques are classically based on document term frequencies, where the terms are obtained after standard text processing, such as text normalization, tokenization, stopping, stemming and named-entity identification. We have been investigating two IR approaches, one based on Okapi term weighting and the other on a Markovian term weighting, combined with query expansion via Blind Relevance Feedback (BRF) using the audio corpus and a parallel text corpus.

As part of the SDR'99 TREC-8 evaluation, 500 hours of unpartitioned, unrestricted American English broadcast data was indexed using both state-of-the-art speech recognizer outputs and manually generated closed captioning .The average word error measured on a representative 10 hour subset of this data was around 20%.

The two IR approaches were shown to yield comparable results [14]. Only small differences in information retrieval performance as given by the mean average precision were observed for automatic and manual transcriptions when the story boundaries are known. These results indicate that the transcription quality may not be the limiting factor on IR performance for current IR techniques.

4. SPOKEN DIALOG SYSTEMS FOR INFORMATION RETRIEVAL

Spoken language systems (SLSs) aim to help a user accomplish a task via interactive dialog. Task and domain knowledge must be used to define the vocabulary and the concepts specific to the application in order to construct appropriate acoustic, language and semantic models. Modelization of spontaneous speech effects, such as hesitations, false starts, and reparations, is particularly important for these systems. In contrast to dictation and transcription tasks where it is relatively straight-forward to select a recognition vocabulary from large written corpora, for spoken language systems there usually are no application-specific training data (acoustic or textual) available. A commonly adopted approach for data collection is to start with an initial system (that may involve a Wizard of Oz configuration) and to collect a set of data which can be used to start an iterative development cycle.

In addition to a speaker-independent, continuous speech recognizer, a spoken language dialog system also includes

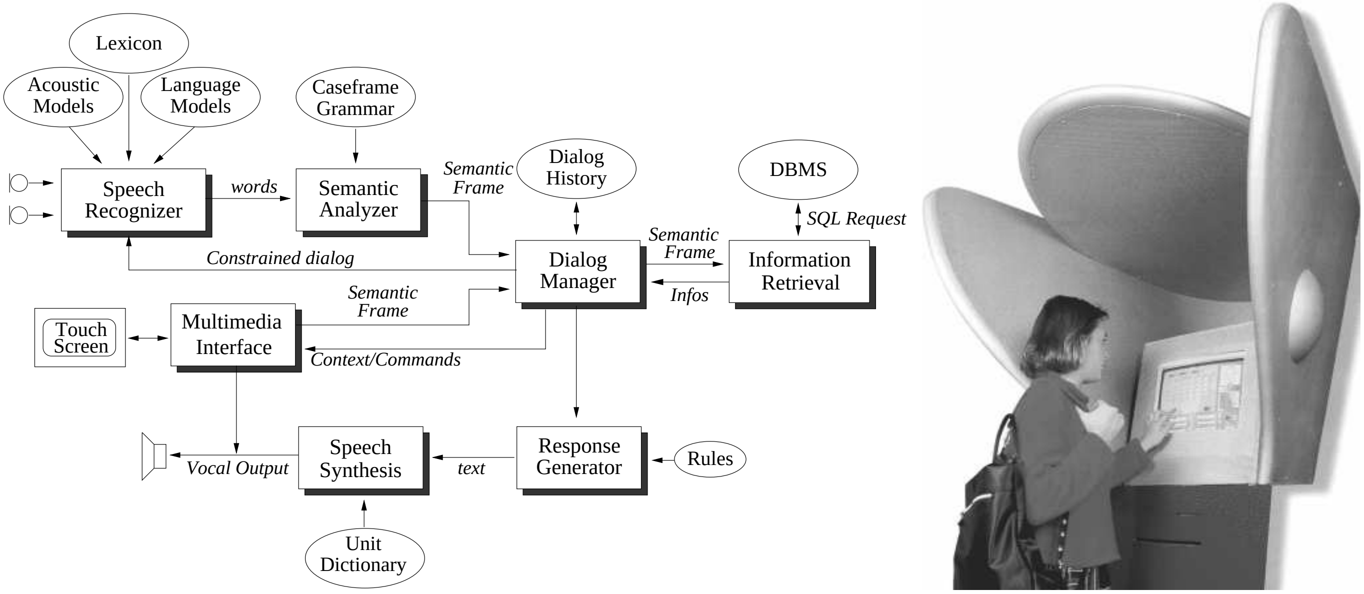

Figure 2: Overview of a spoken language dialog system for information access (left). The MASK kiosk (right).

components for natural spoken language understanding, dialog management, history management, database access, response generation, and speech synthesis. An overview of the LIMSI SLDS architecture is shown in Figure 2. The speech recognizer transforms the input signal into the most probable word sequence (or optionally a word graph), and forwards it to the rule-based natural language understanding component, which generates a semantic frame. A mixed-initiative dialog manager prompts the user to supply any missing information needed for database access and then generates a database query. The retrieved information is transformed into natural language by the response generator (taking into account the dialog context) and presented to the user. Synthesis by waveform concatenation is used to ensure high quality speech output, where dictionary units are put together according to the generated text string. It is becoming increasing clear that dialog management and response generation play an important role in system design and user satisfaction [25].

At LIMSI prototype systems to provide vocal access to train travel information have been developed in the context of several European projects, ESPRIT-MASK [5] and LE RAILTEL, ARISE [19, 21]. Development of systems for more general tourist information (AUPELF) and control of household appliances (Tide HOME) are also underway. These systems have been tested in field trials with naive users.

5. RECENT RESEARCH PROJECTS

- The IDEAL project (CNET 1994-1998) on language identification over the telephone.

  This project explored different approaches to automatic language identification, and evaluated them using a large, multilingual (French, English, German and Spanish) corpus of telephone speech designed for the task [15, 23, 24].

- ESPRIT MASK *Multimodal-multimedia Automated Service Kiosk project* (1994-1998) (http://www.limsi.fr/tlp/mask.html)

  The aim of the MASK project was to explore the use of a spoken language understanding system as part of an advanced user interface for a public service kiosk. A prototype information kiosk, designed to enable the coordinated use of multimodal inputs (speech and touch) and multimedia output (sound, video, text, and graphics) was developed and evaluated in the St. Lazare train station in Paris [5, 16].

- LE-3 ARISE *Automatic Railway Information Systems for Europe* (1996-1999).

  The ARISE project developed speech recognition and spoken dialog technologies that were used in prototype services providing train schedule information [21].

- AUPELF-UREF (1995-1998) projects on “Linguistique, Informatique et Corpus Oraux”.

  Six projects covered the following areas: “Evaluation des systèmes de synthèse,” “Evaluation des systèmes de reconnaissance,” “Evaluation des modèles de langage,” “Evaluation des systèmes de dialogue,” “Corpus de textes,” “Corpus de parole”

- Project under the auspices of the Délégation Générale de l’Armement (DGA) on the Automatic Indexation of Multilingual Broadcasts (1999-2002). This follows a study concerned with the transcription of radio broadcasts and language identification (1997-1998).

- TIDE HOME-AOM project *Home application Optimum Multimedia / multimodal system for Environment control* (1997-2000).
  http://www.swt.iao.fhg.de/home/index.html

  The aim of this project was to develop a single, easy-to-use and coherent usage concept for all household appliances, with one of the communication modes being a natural spoken language interface between the user and system [27].

- The ESPRIT-LTR DISC (1997-1998) and DISC2 (1999) projects *Spoken Language Dialogue Systems and Components Best Practice in Development and Evaluation*
  http://www.disc2.dk/

  The DISC projects aimed at identifying what consitutes best practice in spoken language dialog systems development and evaluation, with the objective of developing reference methodologies [4, 20].

- The LE-4 OLIVE project *A Multilingual Indexing Tool for Broadcast Material Based on Speech Recognition* (1998-2000).
  http://twentyone.tpd.tno.nl/olive/

  The OLIVE project addressed methods to automate the disclosure of the information content of broadcast data thus allowing content-based indexation. Speech recognition was used to produce a time-linked transcript of the audio channel of a broadcast, which was then used to produce a concept index for retrieval.

- The RNRT Vocadis project *Reconnaissance de parole distribuée* (1998-2000).
  http://www.telecom.gouv.fr/rnrt/projets/pvocadis.htm

  This project investigates the concept of ”distributed speech recognition”, which aims to combine the computing power of a central server for speech recognition with local acoustic parameter estimation to ensure high quality analysis with low cost communication.

- Language Engineering LE-5 ALERT *Alert system for selective dissemination* project (2000-2002).
  http://www.fb9-ti.uni-duisburg.de/alert

  The ALERT project aims to associate state-of-the-art speech recognition with audio and video segmentation and automatic topic indexing to develop an automatic media monitoring demonstrator and evaluate it in the context of real world applications. The targeted languages are French, German and Portuguese.

- The RNRT Theoreme project *Thématisation par reconnaissance vocale des médias* (1999-2001).
  http://www.telecom.gouv.fr/rnrt/index_net.htm
  (project 97)

  This project address automated topic detection in audio data.

- LE-5 CORETEX *Improving Core Speech Recognition Technology* (2000-2003).
  http://coretex.itc.it/

  This project aims at improving core speech recognition technologies, which are central to most applications involving voice technology. In particular the project addresses the development of generic speech recognition technology and methods to rapidly port technology to new domains and languages with limited supervision, and to produce enriched symbolic speech transcriptions.

- FP5 IST CIWOS *Combined Image and Word Spotting* (2000-2002).

  The goal of the CIWOS project is to develop and integrate robust unrestricted keyword and image spotting algorithms for audio-visual content retrieval from multimedia databases.

- FP5 IST ECHO *European CHronicles On-line* (2000-2002).
  http://pc-erato2.iei.pi.cnr.it/echo

  The ECHO project aims to develop an infrastructure for access to historical films belonging to large national audiovisual archives. The project will integrate state-of-the-art language technologies for indexing, searching and retrieval, cross-language retrieval capabilities and automatic film summary creation.

6. CONCLUSIONS

This paper has described some of the ongoing speech recognition related research activites at LIMSI. Our goal is develop generic technology that can be applied to a variety of recognition tasks. We are currently mainly investigating two classes of applications: transcription and indexation of broadcast audio data and spoken dialog systems to provide natural, user-friendly interfaces for information access.

ACKNOWLEDGEMENTS

This work has been partially financed by the European Commission and the French Ministry of Defense. The authors thank the other group members who have participated in the non-speech recognition aspects of the spoken dialog systems and spoken document retrieval system described here. These are: Laurence Devillers, Jean-Jacques Gangolf, Veronique Gendner, Yannick de Kercadio, Hélène Maynard, Patrick Paroubek and Sophie Rosset.

REFERENCES

[1] M. Adda-Decker, G. Adda, L. Lamel, "Investigating text normalization and pronunciation variants for German broadcast transcription," *Proc. ICSLP'2000*, Beijing, China, October 2000.

[2] L. Chen, L. Lamel, G. Adda and J.L. Gauvain, "Broadcast News Transcription in Mandarin," *Proc. ICSLP'2000*, Beijing, China, October 2000.

[3] C. Corredor-Ardoy, L. Lamel, M. Adda-Decker, J.L. Gauvain, "Multilingual Phone Recognition of Spontaneous Telephone Speech", *Proc. IEEE ICASSP-98*, **I**, pp. 413–416, Seattle, WA, mai 1998.

[4] L. Dybkjaer et al. "The DISC Approach to Development and Evaluation," *Proc. First International Conference on Language Resources and Evaluation, LREC'98*, **I**, pp. 185-189,Granada, Spain, May 1998.

[5] J.L. Gauvain, S.K. Bennacef, L. Devillers, L.F. Lamel, S. Rosset, "The Spoken Language Component of the Mask Kiosk," in *Human Comfort and Security of Information Systems, Advanced Interfaces for the Information Society*, editors K.C. Varghese S. Pfleger, Springer Verlag, 1997, pp 93–103.

[6] J.L. Gauvain, Y. de Kercadio, L.F. Lamel, G. Adda "The LIMSI SDR system for TREC-8," *Proc. of the 8th Text Retrieval Conference TREC-8*, pp. 405-412, Gaithersburg, MD, November 1999.

[7] J.L. Gauvain, L. Lamel, "Fast Decoding for Indexation of Broadcast Data," *Proc. ICSLP'2000*, Beijing, China, October 2000.

[8] J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI 1997 Hub-4E Transcription System", *Proc.* DARPA *Broadcast News Transcription & Understanding Workshop*, pp. 75-79, Landsdowne, VA February 1998.

[9] J.L. Gauvain, L. Lamel, G. Adda, "Partitioning and Transcription of Broadcast News Data," *ICSLP'98*, **5**, pp. 1335-1338, Dec. 1998.

[10] J.L. Gauvain, L. Lamel, G. Adda, "Recent Advances in Transcribing Television and Radio Broadcasts," *Proc. Eurospeech'99*, **2**, pp. 655-658, Budapest, Sept. 1999.

[11] J.L. Gauvain, L. Lamel, and G. Adda, "Transcribing broadcast news for audio and video indexing," Communications of the ACM, 43(2), Feb 2000.

[12] J.L. Gauvain, L. Lamel, G. Adda and M. Jardino, "The LIMSI 1998 Hub-4E Transcription System", *Proc.* DARPA *Broadcast News Workshop*, pp. 99-104, Herndon, VA, February 1999.

[13] J.L. Gauvain, L. F. Lamel, M. Adda-Decker, "Developments in continuous speech dictation using the ARPA WSJ task," *Proc. IEEE-ICASSP*, pp. 65-68, Detroit, May 1995.

[14] J.L. Gauvain, L. Lamel, Y. Kercadio et G. Adda, "Transcription and Indexation of Broadcast Data, *Proc. IEEE ICASSP'00*, Istanbul, June 2000.

[15] L. Lamel, G. Adda, M. Adda-Decker, C. Corredor-Ardoy, J.J. Gangolf, J.L. Gauvain, "A multilingual corpus for language identification," *Proc. First International Conference on Language Resources and Evaluation, LREC'98*, **II**, pp. 1115-1122, Granada, Spain, May 1998.

[16] L. Lamel, S. Bennacef, J.L. Gauvain, H. Dartigues, J.N. Temem, "User Evaluation of the Mask Kiosk," *Proc. ICSLP'98*, **7**, pp. 2875-2878, Sydney, Australie, decembre 1998.

[17] L. Lamel, J.L. Gauvain, "A phone-based approach to non-linguistic speech feature identification," *Computer Speech and Language*, **9**(1):87-103, January 1995.

[18] L. Lamel, J.L. Gauvain, "Speaker Verification over the Telephone," *Speech Communication*, **31**(2-3), pp. 141-154, June 2000.

[19] L. Lamel, J.L. Gauvain, S.K. Bennacef, L. Devillers, S. Foukia, J.J. Gangolf, S. Rosset, " Field Trials of a Telephone Service for Rail Travel Information," *Speech Communication*, **23**, pp. 67–82, October 1997.

[20] L. Lamel, W. Minker, P. Paroubek, "Towards Best Practice in the Development and Evaluation of Speech Recognition Components of a Spoken Language Dialogue System," *Natural Language Engineering* "special issue on Best Practice in Spoken Language Dialogue Systems Engineering", to appear 2000.

[21] L. Lamel, S. Rosset, J.L. Gauvain, S. Bennacef, M. Garnier-Rizet, B. Prouts "The LIMSI ARISE System," *Speech Communication*, **31**(4), pp. 339–354, August 2000.

[22] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, **9**(2), pp. 171-185, 1995.

[23] D. Matrouf, M. Adda-Decker, J.L. Gauvain, and L. Lamel, "Comparing different model configurations for language identification using a phonotactic approach," *Proc. ESCA EuroSpeech'99*, pp. 387-390, Budapest, sep 1999.

[24] D. Matrouf, M. Adda-Decker, L.F. Lamel, J.L. Gauvain, "*Language identification incorporating lexical information*", *Proc. ICSLP-98*, **2**, pp. 181-184, Sydney, Australie, decembre 1998.

[25] S. Rosset, S. Bennacef and L. Lamel, "Design strategies for spoken language dialog systems", *Proc. ESCA Eurospeech'99*, pp. 1535-1538, Budapest, sep 1999.

[26] H. Schwenk, J.L. Gauvain, "Combining Multiple Speech Recognizers using Voting & Language Model Information," *Proc. ICSLP'2000*, Beijing, China, October 2000.

[27] J. Shao, N.E. Tazine, L. Lamel, B. Prouts, and S. Schröter, "An Open System Architecture for a Multimedia and Multimodal User Interface," *Proceedings of the 3rd TIDE Congress*, Helsinki, June 23-25 1998.