Recent Activities in Spoken Language Processing at LIMSI

Jean-Luc Gauvain and Lori Lamel

Spoken Language Processing Group, LIMSI-CNRS, 91403 Orsay, FRANCE
{gauvain,lamel}@limsi.fr http://www.limsi.fr/tlp

Abstract: This paper summarizes recent activities at LIMSI in multilingual speech recognition and its applications. While the main goal of speech recognition is to provide a transcription of the speech signal as a sequence of words, the same basic technology serves as the first step in other application areas, such as in automatic systems for information access and for automatic indexation of audiovisual data.

SPEECH RECOGNITION

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 (4), 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.

The LIMSI word recognizer (makes use of continuous density HMMs for acoustic modeling of contextual phone units and n-gram statistics estimated on training texts (newspaper texts and/or speech transcriptions) for language modeling. Acoustic modeling uses PLP-like cepstral features derived from a Mel frequency spectrum, with sentence-based cepstral normalization. For each task the recognition vocabulary is selected to maximize lexical coverage, thus minimizing errors due to unknown words. Lexical pronunciations are phonemically-based, with language-specific symbols (7). Specific units are used to model non-speech effects such as silence, filler words, and breath noises.

For transcription tasks, word recognition is carried out in multiple decoding passes, where the information between levels is transmitted via word graphs (5). In the first pass a word graph is generated using a small bigram language model, which is then used to constrain the search space in further decoding passes with more accurate acoustic and language models. Unsupervised acoustic model adaptation may optionally be performed to improve performance.

The applicability of speech recognition techniques for different languages is of particular importance in Europe. At LIMSI, speaker-independent, large vocabulary, continuous speech recognition systems for different European languages (French, German and British English) and for American English have been developed (9). For American English word error rates of about 8% can be achieved for unrestricted newspaper texts from unknown speaker. Word errors on a similar task in French are about 11% (1). The widely used *n*-gram language models are reasonably successful for dictation in English, but are less efficient for more highly inflected languages such as French and German. Applications of this technology are mainly in professional domains, where the systems handle a spoken encoding of written language.

TRANSCRIPTION OF RADIO AND TELEVISION BROADCASTS

Recent advances in the underlying speech recognition technology have led to activity using real-world (or "found") speech data in contrast to speech produced with the aim of being recognized by a machine. One particularly exciting area of research is the transcription of television and radio broadcasts for which automatic processing can ease the workload required for indexation and retrieval purposes (2). This task is challenging from the technical view, as the shows contain data of various acoustic and linguistic nature, with abrupt or gradual transitions between segments. The signal typically consists of prepared speech and spontaneous speech, which 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. Acoustic models trained on clean speech are clearly inadequate to process such inhomogeneous data. 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, new dialects, non-native speakers, etc.

In addition to problems associated with speech transcription, the continuous stream of data needs to be partitioned into homogeneous segments in order to achieve acceptable performance (6). From recent research it appears that current word recognition performance is not critically dependent upon the partitioning accuracy and that any reasonable approach that separates speaker turns and major acoustic boundaries is sufficient. In order to address variability

observed in the linguistic properties, the differences in read and spontaneous speech were analyzed. As a result, filler words and breath noise are explicitly represented in the acoustic and language models. Compound words were introduced as a means of modeling reduced pronunciations for common word sequences. The word error for automatic transcription of continuous broadcast news data is around 20%. This performance seems sufficient for potential applications as demonstrated by the Informedia project at CMU. LIMSI is participating in the LE OLIVE project providing the speech technology for automatic indexation of French and German broadcasts. Related technologies such as speaker or gender identification are applied to improve recognition performance.

SPOKEN LANGUAGE 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 components for natural spoken language understanding, dialog management, history management, database access, response generation, and speech synthesis. 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.

At LIMSI prototype systems to provide vocal access to train travel information have been developed in the context of several European projects, ESPRIT-MASK (3) and LE RAILTEL, ARISE (8). Development of systems for more general tourist information (AUPELF) and control of household appliances (Tide HOME) are also underway. These systems are being tested in field trials with naive users.

REFERENCES

1. Adda, G., Adda-Decker, M., Gauvain, J.L. and Lamel, L., "Text Normalization and Speech Recognition in French," *Proceedings of the European Conference on Speech Technology, EuroSpeech*, Rhodes, Greece, 1997.

2. Gauvain, J.L., Adda, G.,Lamel, L., and Adda-Decker, M., "Transcribing broadcast news: The LIMSI Nov96 Hub4 System," *Proceedings of ARPA Spoken Language Technology Workshop*, Chantilly, Virginia, 56–63, 1997.

3. Gauvain, J.L., Bennacef, S.K., Devillers, L., Lamel,L.F. and Rosset, S., "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.

4. Gauvain, J.L., and Lamel, L., "Large Vocabulary Continuous Speech Recognition: from Laboratory Systems towards Real-World Applications," *IEICE Journal* **J79-D-II**, 2005–2021, (1996).

5. Gauvain, J.L., Lamel, L., Adda, G., and Adda-Decker, M., "Speaker-Independent Continuous Speech Dictation," *Speech Communication* **15**, 21–37, (1994).

6. Gauvain, J.L., Lamel, L., Adda, G., and Adda-Decker, M., "Transcription of Broadcast News," *Proceedings of the European Conference on Speech Technology, EuroSpeech*, Rhodes, Greece, 1997.

7. Lamel, L.F. and Adda, G., "On Designing Pronunciation Lexicons for Large Vocabulary, Continuous Speech Recognition," *Proc. ICSLP'96*, Philadelphia, PA (1996)

8. Lamel, L.F., Gauvain, J.L., Bennacef, S.K., Devillers, L., Foukia, S., Gangolf, J.J., Rosset, S., "Field Trials of a Telephone Service for Rail Travel Information," to appear in *Speech Communication* (1998).

9. Young, S., Adda-Decker, M., Aubert, X., Dugast, C., Gauvain, J.L., Kershaw, D.J., Lamel, L., van Leeuwen, D.A., Pye, D., Robinson, A.J., Steeneken, H.J.M., Woodland, P.C., "Multilingual Large Vocabulary Speech Recognition: The European SQALE Project," *Computer Speech and Language* **11**(1), 73-89 (1997).