Automatic Processing of Broadcast Audio in Multiple Languages

Lori Lamel and Jean-Luc Gauvain
Spoken Language Processing Group
LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France
{lamel,gauvain}@limsi.fr   http://www.limsi.fr/tlp

ABSTRACT

This paper addresses recent progress in LVCSR in multiple languages which has enabled the processing of broadcast audio for information access. At LIMSI, broadcast news transcription systems have been developed for seven languages. Automatic processing to access the content must take into account the specificities of audio data, such as needing to deal with the continuous data stream and an imperfect word transcription, and specificities of the language. Some near-term applications are audio data mining, structurization of audiovisual archives, selective dissemination of information and media monitoring.

1. INTRODUCTION

With the rapid expansion of different media sources (via the radio, television, Internet), there is a pressing need for automatic processing of such audio streams [8, 16, 18]. Transcribing and annotating audio data is a necessary step in order to provide access to its content, and large vocabulary continuous speech recognition is a key technology for automatic processing. Broadcast audio is challenging to process as it consists of a continuous flow of audio data comprised of segments with various acoustic and linguistic natures. Processing such inhomogeneous data thus requires appropriate modeling at both levels.

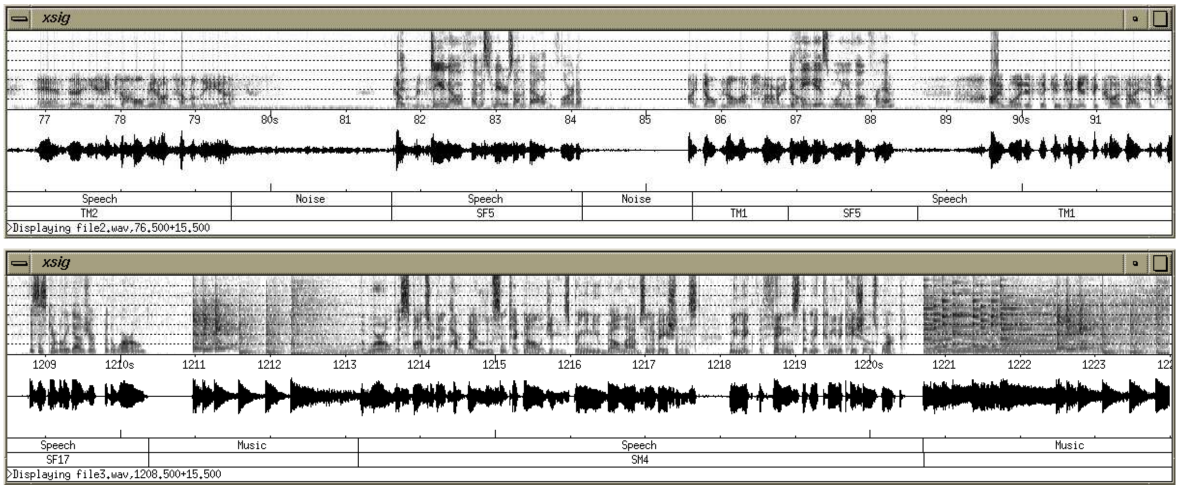
Since most of the linguistic information is encoded in the audio channel of video data, once transcribed it can be accessed using text-based tools. The annotations can be used for indexation and retrieval purposes, as explored in the EC OLIVE [4] and ECHO [3] projects for the disclosure of audiovisual archives. Spoken document retrieval supports random access to relevant portions of audio documents, reducing the time needed to identify recordings in large multimedia databases [9]. Another application area concerns detecting and tracking events on particular subjects of interest. In the EC ALERT [1] project and the French RNRT THEOREME [5] project, the aim is to combine state-of-the-art speech recognition with audio and video segmentation and automatic topic indexing to develop a demonstrator for selective dissemination of information, and to evaluate it within the context of real-world applications. Versions of the LIMSI broadcast news transcription system have been developed for the American English, French, German, Mandarin, Portuguese, Spanish and Arabic languages.

2. TRANSCRIBING BROADCAST AUDIO

The ability of systems to deal with non-homogeneous data as is found in broadcast audio (changing speakers, languages, backgrounds, topics) has been enabled by advances in a variety of areas including techniques for robust signal processing 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 much larger vocabularies than in the past - 64 k words or more is common to reduce errors due to out-of-vocabulary words.

The LIMSI broadcast news (BN) transcription system for automatic indexation has two main components: an audio partitioner and a speech recognizer. The goal of audio partitioning is to divide the acoustic signal into homogeneous segments, labeling and structuring the acoustic content of the data, and identifying and removing non-speech segments. The BN audio partitioner relies on an audio stream mixture model [12]. While it is possible to transcribe the continuous stream of audio data without any prior segmentation, partitioning offers several advantages over this straightforward solution. Additional nonlinguistic information can be extracted from the audio signal, such as the segmentation into speaker turns, the speaker identities, and background acoustic conditions. This information can be used both directly and indirectly for indexation and retrieval purposes. 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. Eliminating non-speech segments also substantially reduces the computation time. The partitioning process produces a set of non-overlapping speech segments, which usually correspond to speaker turns with speaker, gender and telephone/wide-band labels (see Fig. 1).

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. The LIMSI continuous speech recognizer makes use of 4-gram statistics for language modeling and of continuous density hidden Markov models (HMMs) with Gaussian mixtures for acoustic modeling. Each word is represented by one or more sequences of context-dependent phone models as determined by its pronunciation. The acoustic and



>Displaying file2.wav,76.500+15.500

>Displaying file3.wav,1208.500+15.500

Figure 1: Spectrograms illustrating data partitioning results. The automatically generated segment type is one of: Speech, Music, or Noise. For the speech segments the cluster labels specify the identified bandwidth (T=telephone-band/S=wideband) and gender (M=male/F=female), and cluster number.

language models are trained on large, representative corpora for each task and language. Word recognition is performed in three steps: 1) initial hypothesis generation, 2) word graph generation, 3) final hypothesis generation. The first step, generates initial hypotheses which are used for cluster-based acoustic model adaptation of both the means and variances using the MLLR technique. Acoustic model adaptation is quite important for reducing the word error rate, with gains on the order of 20%. Experiments indicate that the word error rate of the first pass is not critical for adaptation. The second decoding step generates word graphs which are used in the third pass with a 4-gram language model and readapted acoustic models to generate the final hypothesis. The 4-gram single pass dynamic network decoder is described in detail in [11], along with some measures of recognition performance as a function of decoding strategy. Using this decoder, unrestricted BN data can be decoded in less than 1xRT (including partitioning) with a word error rate under 30%. The same decoding strategy has been successively applied to the BN transcription in other languages with somewhat comparable word error rates.

3. MULTILINGUALITY

A characteristic of the broadcast news domain is that, at least for what concerns major news events, similar topics are simultaneously covered in different emissions and in different countries and languages. Automatic processing of contemporaneous data sources in different languages can serve for multi-lingual indexation and retrieval. Multilinguality is of particular interest for media watch applications, where news may first break in another country or language.

Porting a recognizer to another language necessitates modifying those the system components which incorporate language-dependent knowledge sources such as the phone set, the recognition lexicon, phonological rules and the language model. Other considerations are the acoustic confusability of the words in the language (such as homophone, monophone, and compound word rates) and the word coverage of a given size recognition vocabulary. There are two predominant approaches taken to bootstrapping the acoustic models for another language. The first is to use acoustic models from an existing recognizer and a pronunciation dictionary to segment manually annotated training data for the target language. If recognizers for several languages are available, the seed models can be selected by taking the closest model in one of the available language-specific sets. An alternative approach is to use a set of global acoustic models, that cover a wide number of phonemes [17]. This approach offers the advantage of being able to use the multilingual acoustic models to provide additional training data, which is particularly interesting when only very limited amounts of data ($<$ 10 hours) for the target language are available.

At LIMSI broadcast news transcription systems have been developed for the American English, French, German, Mandarin, Spanish, Arabic and Portuguese languages. To give an idea of the resources used in developing these systems, the training material are shown in Table 1. It can be seen that there is a wide disparity in the available language resources for a broadcast news transcription task: for American English, 200 hours of manually transcribed acoustic training are available from the LDC, compared with only about 20-50 hours for the other languages. Obtaining appropriate language model training data is even more difficult. While newspaper and newswire texts are becoming widely available in many languages, these texts are quite different than transcriptions of spoken language. Over 10k hours of commercial transcripts are available for American English, and many TV stations provide closed captions. Such data are not available for most other languages, and in some countries it is illegal to sell transcripts.

Some of the system characteristics are shown in Table 2, along with indicative recognition performance rates. The

	Audio			*Text (words)*	
Language	*Radio-TV sources*	*Duration*	*Size*	*News*	*Com.Trans.*
English	ABC, CNN, CSPAN, NPR, PRI, VOA	200h	1.9M	790M	240M
French	Arte, TF1, A2, France-Info, France-Inter	50h	0.8M	300M	20M
German	Arte	20h	0.2M	260M	-
Mandarin	VOA, CCTV, KAZN	20h	0.7M(c)	200M(c)	-
Portuguese	9 sources	3.5h	∼35k	70M	-
Spanish	Televisa, Univision, VOA	30h	0.33M	295M	-
Arabic	tv: Aljazeera, Syria; radio: Orient, Elsharq, ...	50h	0.32M	200M	-

Table 1: Approximate sizes of the transcribed audio data and text corpora used for estimating acoustic and language models. For the text data, newspaper texts (News) and commercial transcriptions (Com.Trans.). The American English, Spanish and Mandarin data are distributed by the LDC. The German data come from the EC Olive project and the French data from Olive and from the DGA. The Portuguese data are part of the pilot corpus used in the EC Alert project. The Arabic data were produced by the Vecsys company in collaboration with the DGA.

	Lexicon			*Language Model*		*Test*	
Language	*#phones*	*size (words)*	*coverage*	*N-gram*	*ppx*	*Duration*	*%Werr*
English	48	65k	99.4%	11M fg, 14M tg, 7M bg	140	3.0h	20
French	37	65k	98.8%	10M fg, 13M tg, 14M bg	98	3.0h	23
German	51	65k	96.5%	10M fg, 14M tg, 8M bg	213	2.0h	25(n)-35(d)
Mandarin	39	40k+5k(c)	99.7%	19M fg, 11M tg, 3M bg	190	1.5h	20
Spanish	27	65k	94.3%	8M fg, 7M tg, 2M bg	159	1.0h	20
Portuguese	39	65k	94.0%	9M tg, 3M bg	154	1.5h	40
Arabic	40	60k	90.5%	11M tg, 6M bg	160	5.7h	20

Table 2: Some language characteristics. For each language are specified: the number of phones used to represent lexical pronunciations, the approximate vocabulary size in words (characters for Mandarin) and lexical coverage (of the test data), the language model size, the test data perplexity, duration and the word/character error rates. For Arabic the vocabulary and language model are vowelized, however the word error rate does not include vowel or gemination errors. For German, separate word error rates are given for broadcast news (n) and documentaries (d).

word error rate on unrestricted American English broadcast news data is about 20% [10, 11]. The transcription systems for French and German have comparable error rates for news broadcasts [6]. The character error rate for Mandarin is also about 20% [7]. 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, documentaries are particularly challenging to transcribe, as the audio quality is often not very high, and there is a large proportion of voice over.

4. ACCESSING CONTENT

The automatically generated partition and word transcription can be used for indexation and information retrieval purposes. Techniques commonly applied to automatic text indexation can be applied to the automatic transcriptions of the broadcast news radio and TV documents. Three of the applications investigated are Spoken Document Retrieval (SDR) [10], Topic Detection and Tracking (TDT) [15] and Selective Dissemination of Information [1].

The spoken document retrieval system was evaluated in the TREC SDR track, with known story boundaries. In order to assess the effect of the recognition time on the information retrieval results test corpus of 557 hours of broadcast news data was transcribed using two decoder configurations: a single pass 1.4xRT system and a three pass 10xRT system. The word error rate measured on a 10h test subset [9] were 32.6% and 20.5% respectively. With query expansion and blind relevance feedback comparable IR results in terms of mean average precision (the measure used in the TREC benchmarks) are obtained using the closed captions (54.3%) and the 10xRT transcriptions (53.9%) and a moderate degradation (4% absolute) is observed using the 1.4xRT transcriptions. When stories boundaries are not known a priori a double windowing technique [10] is proposed to locate stories. The IR result with automatically detected story boundaries is 52.3% which can be compared with 59.6% if manual segmentations (provided by NIST) are used. It should also be noted that the manual segmentation removes all non-story segments such as advertising. This reduces the risk of having out-of-topic hits and explains part of the difference between the results.

Topic tracking consists of identifying and flagging on-topic segments in a data stream. These techniques can be applied in media-watch applications [1] as well as used to structure multimedia digital libraries [3]. A topic tracking system was developed at Limsi [15] which relies on a unigram topic model, where a topic is defined by a set of topic related audio and/or textual documents. These documents were used to train a topic model, which was used to locate on-topic documents in an incoming stream.

5. PORTABILITY AND ADAPTABILITY

With today's technology, the adaptation of a recognition system to a different task or language requires the availability of sufficient amounts of transcribed training data. Obtaining such data is usually an expensive process in terms of manpower and time. Recent work has focused on reducing this development cost [2]. One approach is to use existing recognizer components (developed for other tasks or languages) to automatically transcribe task-specific training data. Although in the beginning the error rate on new data is likely to be rather high, this speech data can be used to retrain a recognition system. If carried out in an iterative manner, the speech corpus can be cumulatively extended over time *without* direct manual transcription. This approach has been investigated in [13, 14, 19]. In [14] it was found that somewhat comparable acoustic models could be estimated on 400 hours automatically annotated data from the TDT-2 corpus and 150 hours of carefully annotated data.

The same basic idea was used to develop acoustic models for the Portuguese language for which substantially less manually transcribed data are available. Initial acoustic model trained on the 3.5 hours of available data were used to transcribe 30 hours of Portuguese TV broadcasts. These acoustic models had a word error rate of 42.6%. By training on the 30 hours of data using the automatic transcripts the word error was reduced to 39.1%. This preliminary experiment supports the feasibility of lightly supervised and unsupervised acoustic model training.

6. CONCLUSIONS

Automatic speech recognition is a key technology for audio and video indexing. Most of the linguistic information is encoded in the audio channel of video data, which once transcribed can be accessed using text-based tools. This is in contrast to the image data for which no common description language is widely adopted. It appears that the word error rates obtained with state-of-the-art systems (on the order of 20%), are sufficient to enable a variety of near-term applications such as audio data mining, selective dissemination of information (News-on-Demand), media monitoring, content-based audio and video retrieval.

While this paper has drawn examples from ongoing research activities at LIMSI in automatic transcription and indexation of broadcast data, this is a research area with wide international activity. Much of the research, which is at the forefront of todays technology, aims at addressing our partners' needs for advanced audio processing technologies.

Given the reliance of todays most performant systems on large training corpora, porting across languages or domains first requires obtaining the necessary resources. Research is underway to reduce the need for manually annotated training data, thus reducing the human investment needed for system development. However, obtaining a pronunciation lexicon still substantial manual effort in order to represent spoken language, and in particular to deal with foreign words and proper names which are common in broadcast data. Our experience is that while the same basic technologies and development strategies appear to port from one language to another, to obtain optimal performance language specificities must be taken into account.

7. ACKNOWLEDGMENTS

This work has been partially financed by the European Commission and the French Ministry of Defense. The authors thank their colleagues in the Spoken Language Processing group at LIMSI for their participation in the development of different aspects of the automatic transcription and indexation systems reported here.

REFERENCES

[1] http://www.fb9-ti.uni-duisburg.de/alert

[2] http://coretex.itc.it

[3] http://pc-erato2.iei.pi.cnr.it/echo/

[4] http://twentyone.tpd.tno.nl/olive

[5] ttp://www-mrim.imag.fr/projets/theoreme.php

[6] M. Adda-Decker et al.,"Investigating text normalization and pronunciation variants for German broadcast transcription," *ICSLP'2000*.

[7] L. Chen et al.,"Broadcast News Transcription in Mandarin," *ICSLP'2000*.

[8] C. Djeraba, ed., Special issue on "Content-Based Multimedia Indexing and Retrieval," *Multimedia Tools & Applications*, **14**(2), June 2001.

[9] J.S. Garofolo et al., "The TREC Spoken Document Retrieval Track: A Success Story," *RIAO'00*, April 2000.

[10] J.L. Gauvain et al., "The LIMSI SDR System for TREC-9", *TREC-9*, Nov 2000.

[11] J.L. Gauvain, L. Lamel, "Fast Decoding for Indexation of Broadcast Data," *ICSLP'2000*.

[12] J.L. Gauvain et al., "Partitioning and Transcription of Broadcast News Data," *ICSLP'98*.

[13] T. Kemp, A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments," *Eurospeech'99*.

[14] L. Lamel et al., "Lightly Supervised and Unsupervised Acoustic Model Training," *Computer, Speech & Language*, Jan. 2002.

[15] Y.Y. Lo, J.L. Gauvain "The LIMSI Topic Tracking System for TDT2001," *Topic Detection & Tracking Workshop*, Nov 2001.

[16] M. Maybury, ed., Special Section on "News on Demand", *Communications of the ACM*, 43(2), Feb 2000.

[17] T. Schultz, A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, **35** (1-2), Aug. 2001.

[18] M. Yuschik, ed. 'Special issue on "Multimedia Technologies, Applicatins and Perfomrance," *International Journal of Speech Technology* **4** (3/4), July-Oct 2001.

[19] G. Zavaliagkos, T. Colthurst, "Utilizing Untranscribed Training Data to Improve Performance," *DARPA Broadcast News Transcription & Understanding Wshop*, 1998.