SPEECH TRANSCRIPTION IN MULTIPLE LANGUAGES

L. Lamel, J.L. Gauvain, G. Adda, M. Adda-Decker, L. Canseco,
L. Chen, O. Galibert, A. Messaoudi, H. Schwenk

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

ABSTRACT

This paper summarizes recent work underway at LIMSI on speech-to-text transcription in multiple languages. The research has been oriented towards the processing of broadcast audio and conversational speech for information access. Broadcast news transcription systems have been developed for seven languages and it is planned to address several other languages in the near term. Research on conversational speech has mainly focused on the English language, with initial work on the French, Arabic and Spanish languages. Automatic processing must take into account the characteristics of the audio data, such as needing to deal with the continuous data stream, specificities of the language and the use of an imperfect word transcription for accessing the information content. Our experience thus far indicates that at today's word error rates, the techniques used in one language can be successfully ported to other languages, and most of the language specificities concern lexical and pronunciation modeling.

1. INTRODUCTION

Much of the recent developments in speech-to-text transcription have been oriented towards the transcription of broadcast data, of telephone conversations, and most recently, multi-person meetings. Some near-term applications of the technology are audio data mining, structurization of audio and audiovisual archives, selective dissemination of information, media monitoring, and surveillance. 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. These audio data sources are challenging to process as they consists of a continuous flow of audio data comprised of segments with various acoustic and linguistic natures, from a variety of talkers and may included speech in multiple languages. 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. 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 [8].

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 partitioning process relies on an audio stream mixture model [4] and produces a set of non-overlapping speech segments, which usually correspond to speaker turns with speaker, gender and telephone/wide-band labels.

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 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 language models are trained on large, representative corpora for each task and language. Word recognition is performed in multiple steps. The first step generates initial hypotheses with a 3-gram language model (LM) 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 er-

0-7803-8484-9/04/$20.00 ©2004 IEEE

ror rate of the first pass is not critical for adaptation. Then word lattices are generated using a 2-gram LM and rescored with a 4-gram LM after conversion to a consensus network.

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 for which comparable language resources are available. Versions of the LIMSI broadcast news transcription system have been developed for the American English, French, German, Mandarin, Portuguese, Spanish and Arabic languages, and work on the Dutch and Italian languages is underway.

3. CONVERSATIONAL SPEECH

It is well known that transcribing conversational telephone speech poses many challenges [1]. Challenges at the acoustic level for this type of data concern speaker normalization, the need to cope with channel variability, and the need for efficient speaker adaptation with small amounts of adaptation data. On the linguistic side the primary challenge is to cope with the limited amount of language model training data.

There are notable differences in the speaking styles of conversational telephone speech (CTS) and broadcast news. Broadcast speech is much closer to written language than is conversational speech, where different conventions are observed. The conversational speech may have quite varied acoustic conditions which the quality being affected by the telephone handset, the background noise (other conversations, music, street noise, etc), as well as a much higher proportion of interruptions, overlapping speech and third person interjections or side conversations. In terms of linguistic content, there are many more speech fragments, hesitations, restarts and repairs, as well as backchannel confirmations to let each interlocutor know the other person is listening than in BN data. The first person singular form is much more predominant in conversational speech than in BN. Another major difference from BN is that some interjections such as “uh-huh” and “mhm” (meaning yes) and “uh-uh” (meaning no) that are considered as non-lexical items in BN, need to be recognized since they provide feedback in conversations and help maintain contact. The word “uhhuh”, which serves both to signal agreement and a backchannel “I’m listening”, accounts for almost 1% of the running words in the CTS data. The most common word in the English CTS data, “I”, accounts for almost 4% of all word occurrences, but only about 1% of the word occurrences in BN. Similar observations can be made for the French language, for which we have recently started developing a CTS system. There are a large number of verb forms in the CTS data that were rarely observed in the French BREF and BN data. This less formal conversational speaking style has a much larger proportion of the first and second person forms (“je” and “tu”).

The LIMSI SWB speech-to-text system [5] relies on the same basic components as the LIMSI BN system [4]. Additional features specific to the SWB system are: vocal-tract length normalization, multiple regression class MLLR adaptation, pronunciation probabilities, neural-network language model, and consensus decoding. Some of these techniques (in particular Vocal Tract Length Normalization (VTLN) and pronunciation probabilities) which had not helped in our BN transcription system, quite significantly improve the performance of our SWB system.

Whereas for the BN task it is relatively easy to find a variety of related texts that can be processed and used as training materials, for conversational speech, the only available resource is the transcripts of the audio data. To deal with this problem we have tried to select “conversational style” texts from other sources, such as BN data, to provide additional training data, and have used LM smoothing via a neural network [5].

Most of our experience on the transcription of conversational speech has been for American English. A notable difference in French is the use of slang and “verlan”[1] in French. Certain strong word reductions have an accepted written form in English such as “he’d, gonna, dunno, gotta”, whereas in French there is no common written form for “chais pas” for “je ne sait pas” (*I don’t know*) or “ste” in the place of “cet” (*this*). For the moment we have chosen to include some of the colloquial word forms in the recognition word list, but have not yet determined a satisfying manner to handle the strong reductions other than to add compound words. The Arabic language is even more challenging in that many conversations are carried out in dialects for which there is no standard written form. We expect that many more differences from broadcast speech and across languages will arise as our work progresses.

4. MULTILINGUALITY

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 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. Two predominant approaches are taken to bootstrap acoustic models for another language. The first uses acoustic models from an

[1] Verlan refers to speaking by reversing the syllables in the word, similar to pig Latin in English. So the words “bizarre” becomes “zarbi”, and “parent” is “rentpa”

	Audio			*Text (words)*	
Language	*Radio-TV sources*	*Duration*	*Size*	*News*	*Com.Trans.*
English	ABC, CNN, CSPAN, NPR, PRI, VOA	200h(+400h)	1.9M	1B	242M
French	Arte, TF1, A2, France-Info, France-Inter	50h	0.8M	320M	21M
German	Arte	30h	0.2M	315M	-
Mandarin	VOA, CCTV, KAZN, CNR, CBS, CTS	20h(+120h)	0.7M(c)	200M(c)	10.2M
Portuguese	9 sources	3.5h(+30h)	~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 English, Mandarin and Portuguese the amount of untranscribed audio data used for lightly supervised acoustic model training is also given in parentheses. 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 and Alert projects and the French data from Olive, Alert 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		*Test*	
Language	*#phones*	*size (words)*	*coverage*	*Duration*	*%WER/CER*
English	48	65k	99.4%	3.0h	12
French	37	65k	98.8%	3.0h	18
German	49	300k	98.6%	2.0h	18
Mandarin	61	40k+5k(c)	99.7%	1.5h	20
Spanish	27	65k	94.3%	1.0h	20
Portuguese	39	65k	94.0%	1.5h	40
Arabic	40	60k	90.5%	5.7h	20

Table 2: Some language characteristics. For each language 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), duration and the indicative word/character error rates are given.

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, covering a wide number of phonemes [10]. This approach offers the advantage of being able to use the multilingual acoustic models to provide additional training data, which is of particular interest when only very limited amounts of data ($<$ 10 hours) for the target language are available.

For some languages it is relatively straightforward to generate a pronunciation lexicon using grapheme-to-phoneme rules. This is the case for languages such as French, Spanish, Portuguese, Italian and Arabic (if, as in our case, the orthographic form includes vowels). However, even if grapheme-to-phoneme conversion is a viable solution, there are a number of words that need to be treated separately, such as proper names and borrowed words from other languages for which the standard pronunciation rules do not apply. In the CoreTex project [2] data-driven and rule-based methods for automatic pronunciation generation were investigated.

Broadcast news transcription systems have been developed for the American English, French, German, Mandarin, Spanish, Arabic and Portuguese languages. Table 1 gives an idea of the resources used in developing these systems. 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. While newspaper and newswire texts are becoming widely available in many languages, commercial transcripts or closed-captions for are much more difficult to obtain. Over 10k hours of commercial transcripts are available for American English, and many TV stations provide closed captions. Such data are often not available for 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 word error rate on unrestricted American English broadcast news data is about 12%. The transcription systems for French, German and Spanish have comparable error rates for news broadcasts [9]. The use of a 300k word lexicon for German reduces the OOV rate to 1.5%, from about 5% with 65k words. The larger vocabulary combined with additional language model training texts reduced the word error rate from 22.2% to 18.9% on 6 news broadcast shows from Tagesschau [9]. The character error rate for Mandarin is also about 20% [3]. Because Mandarin Chinese is a character-based language, any Chinese text can be covered by a character string. This means that if all individual characters are included in the recognition lexicon, there is no problem of out-of-vocabulary items. The Mandarin phone set has 24 consonants and 11 vowels, each having possible 3 tones. This is a simplified representation of tone where the 5 tones for vowels are collapsed into three: flat (tones 1 and 5), rising (tones 2 and 3), and falling (tone 4).

The Arabic language poses challenges somewhat different from the other languages (mostly Indo-European Germanic or Romance) we have worked with. It is a strongly consonantal language with nominally only six vowels, three long and three short. An agglutinative language, it has many different word forms for a given root, produced by appending articles at the word beginning and possessives at the word end. Arabic is also written and read from right to left, which requires modification to the text processing utilities. Texts are typically non-vowelized, meaning the the short vowels are not indicated, and there are typically several possible (generally semantically linked) vowelizations for a given written word. The audio data used in the Arabic system were transcribed with vowels, allowing these to be explicitly modeled in the lexicon and acoustically. Although a vowelized representation is used in the lexicon and language model, the word error rate does not include vowel or gemination errors. (This rate can as much as double if such errors are counted.)

Obtaining sufficient amounts of transcribed audio training data is usually expensive in terms of both 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 re-train 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 [6, 7, 12]. In [7] it was found that somewhat comparable acoustic models could be estimated on 400 hours of automatically annotated data from the TDT-2 corpus and 150 hours of carefully annotated data. The same basic idea was used to exploit the TDT data (in Mandarin and Arabic) to improve the acoustic models for these languages [11].

5. CONCLUSIONS

Automatic speech recognition is a key technology for audio and video indexing. It appears that the word error rates obtained with state-of-the-art systems (on the order or below 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. 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.

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. Given the reliance of today's 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. We expect that as word error rates are lowered, the language dependent issues should become more important, and language specific knowledge will help to improve performance.

REFERENCES

[1] C.S. Culhane, B.J. Wheatley, "Hub-5 Conversational Speech Recognition," DARPA *Speech Recognition Workshop,* Feb. 1996.

[2] CORETEX: Improving Core Speech Recognition Technology http://coretex.itc.it

[3] L. Chen et al.,"Transcribing Mandarin Broadcast News," *IEEE ASRU'2003.*

[4] J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI Broadcast News Transcription System" *Speech Communication* 37(1-2), 2002.

[5] J.L. Gauvain et al., "Conversational Telephone Speech Recognition," *ICASSP'03.*

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

[7] L. Lamel, J.L. Gauvain, G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," *Computer, Speech & Language*, **16**(1), 2002.

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

[9] K. McTait, M. Adda-Decker, "The 300k LIMSI German Broadcast News Transcription System," *Eurospeech'03*

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

[11] R. Schwartz et al., "Speech Recognition in Multiple Languages and Domains:The 2003 BBN/LIMSI EARS Sys tem," *ICASSP'04.*

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