Speech Processing for Audio Indexing*

Lori Lamel and Jean-Luc Gauvain

LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
{lamel,gauvain}@limsi.fr

Abstract. This paper addresses some of the recent trends in speech processing, with a focus on speech-to-text transcription as a means to facilitate access to multimedia information in a multilingual context. A brief overview of automatic speech recognition is given along with indicative performance measures for a range of tasks. Enriched transcriptions, that is enhancing the automatic word transcripts with meta-data derived from the audio data is discussed, followed by some hightlights of recent progress and remaining challenges in speech recognition.

1 Introduction

The last decade has witnessed major advances in spoken language technologies, with a growing interest in applications that rely on techniques for automatic structurization of multimedia, multilingual data. Although the different media types typically bring complementary information, for most documents much of the accessible content is provided by the audio and text streams. Thus speech and language processing technologies are key components for indexing. Some of the applications that can potentially make use of spoken language technologies are the creation and access to digital multimedia libraries, media monitoring services to provide selective dissemination of information based on automatic detection of topics of interest, and more generally speaking as News on Demand and Internet watch services which already are available for text documents. Developing speech technologies is by nature an interdisciplinary process, requiring knowledge and competence in a range of disciplines including signal processing, acoustics, phonetics, linguistics, artificial intelligence, etc. In addition to speech transcription, speech processing techniques can be used to provide other metadata, such as the language being spoken, the identity of the speaker, as well as to locate named entities or identify topics.

While the performance of speech recognition technology has dramatically improved for a number of 'dominant' languages (English, Mandarin, Arabic, French, Spanish, ...), generally speaking technologies for language and speech processing are available only for a small proportion of the world's languages. By several estimations there are over 6000 spoken languages in the world, but only about 15% of them also are written. Text corpora, which can be useful for training the language models used by speech recognizers, are becoming more and more readily available on the Internet. The site

* This work has been partially financed under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022 and by OSEO under the Quaero program.

A. Ranta, B. Nordström (Eds.): GoTAL 2008, LNAI 5221, pp. 4–15, 2008.

http://www.omniglot.com lists about 800 languages that have a written form. According to http://www.nvtc.gov/lotw the top 10 languages on the Internet account over 80% of use, with the dominant language being English (almost 30%) and the second Chinese (14%). For speech recognition training purposes the best texts are speech transcripts, or texts that are close to spoken language. For prepared speech, such as broadcast news type data, newspaper texts are quite useful, and some efforts have been made to transform such material to better match spoken language [12]. For more conversational speech less formal texts are more appropriate and there have been recent effort to locate such data from the web, for example, from blogs [9].

It is difficult to estimate the amount of audio data on the Internet. A study by the University of Berkeley School of Information Management and Systems[1], attempts to estimate the proportions of different data types based on the file types and sizes. From these estimations of file size, about 30% of the files correspond to text data, about 20% image, 5% video and 3% audio. Considering worldwide sources of radio and television, about 100 million hours of original programming (about 20% from the US) are broadcast per year, representing about 10 terawords of data.

There have been numerous national and international projects addressing different aspects of processing multimedia, multilingual data for information access. Perhaps the longest running project is the National Science Foundation (NSF) Digital Libraries Informedia project (http://www.informedia.cs.cmu.edu), which started in the mid 1990s, aims to incorporate automatic text, speech, image and video processing to enable content-based search in multimedia digital archives. A list of ongoing national and European sponsored projects can be found on the web site of the Chorus coordinating action (http://www.ist-chorus.org/projects.asp), some of which include research on speech and audio processing.

During the last twenty years there has also been an accompanying growth in a support infrastructure for data collection, annotation and evaluation. Concerning data, the most notable actors are the Linguistic data consortium (LDC, http://www.ldc.upenn.edu), founded in 1992 with the goal of developing a mechanism for the creation of and the widespread sharing of linguistic resources for linguistic research, and the European Language Resources Association (ELRA, http://www.elra.info), founded in 1995 with the aim of promoting language resources and evaluation for the Human Language Technology sector. The Speech Group at the National Institute of Standards and Technology (NIST) has been organizing benchmark evaluations for a range of human language technologies (speech recognition, speaker and language recognition, spoken document retrieval, topic detection and tracking, automatic content extraction, spoken term detection) for over 20 years, recently extending to related multi-modal technologies[2]. Comparative evaluation of technologies in international campaigns is important in order to objectively assess the methods and models developed, and serves to increase the information exchange among participants. These evaluations require the development of methods and metrics to measure performance, as well as the annotation of

1 http://www2.sims.berkeley.edu/research/projects/how-much-info-2003

2 See http://www.nist.gov/speech/tests for a summary of previous and current evaluation campaigns.

development and test data. The post-evaluation workshops provide the opportunity for each participant to describe their research and development work in preparation for the evaluation, thus promoting the exchange of information. The most promising techniques are seen to be quickly adopted by other members of the research community, thus leading to rapid advances in the state-of-the-art.

Many of the recent advances can be attributed to the increased use of real world data, with its challenges and advantages. There has been a shift towards algorithms that can benefit from large corpora, and the development of methods to reduce the amount of supervision required for model training. While this paper focuses on speech recognition, there has been a trend to use corpus-based methods for other technologies, such as speech synthesis, speech understanding and machine translation of speech.

2 Speech Recognition Basics

Most state-of-the-art automatic speech recognition systems make use of statistical models, the principles of which have been known for many years [7,14]. From this point of view, speech is assumed to be generated by a language model which provides estimates of $\Pr(w)$ for all word strings w, and an acoustic model encoding the message w in the signal x, which is represented by a probability density function $f(x|w)$. Given the observed acoustic signal, the goal of speech recognition is to determine the most likely word sequence. The speech decoding problem thus consists of maximizing the probability of the word sequence w given the speech signal x, or equivalently, maximizing the product $\Pr(w)f(x|w)$. Considerable progress has been made in recent years in part due to the availability of large speech and text corpora, along with increased processing power which have allowed more complex models and algorithms to be implemented. The advances in acoustic, language and pronunciation modeling have enabled reasonable performance to be obtained for a range of data types and acoustic conditions.

The principle problems in speech recognition have been the focus of many years of research. The variability observed in the acoustic signal is due to multiple factors, including the linguistic message and the characteristics of the speaker, acoustic environment, recording conditions and transmission channel. Figure 1 shows the main components of a speech recognition system using statistical methods for training and decoding [19]. The main knowledge sources are the speech and text training data and the pronunciation lexicon. Acoustic and language model training relies on the preprocessing and normalization of the data. In general, speech data is manually transcribed, however recent research has been directed at reducing the need for supervision. Concerning the text corpus, after some initial processing to remove material unsuitable for sentence-based language modeling, such as tables and lists, the texts need to be normalized. This step, which helps reduce lexical variability and transforms the texts to better represent spoken language, is typically language specific. It includes rules to the process numbers, abbreviations and acronyms, and may also concern how hyphenated words, other compounds or words with apostrophes are treated.

The most popular language models for large vocabulary speech recognition [27] are *n*-gram models, which attempt to capture the syntactic and semantic constraints by estimating the frequencies of sequences of *n* words. The probability of a given

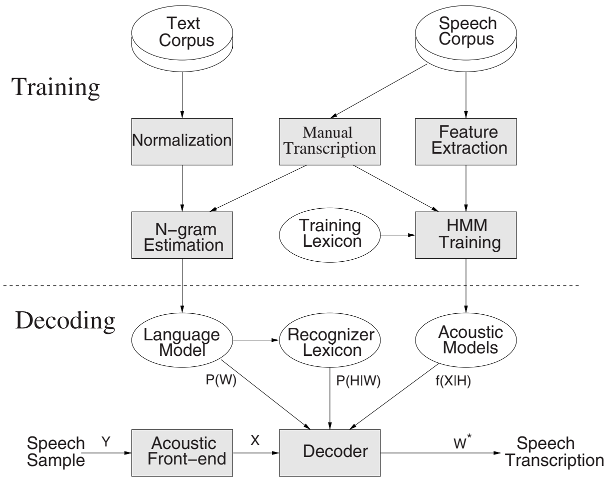

Fig. 1. System diagram of a speech recognizer based on statistical models, including training and decoding processes

word string $(w_1, w_2, ..., w_k)$ is approximated by $\prod_{i=1}^{k} \Pr(w_i | w_{i-n+1}, ..., w_{i-2}, w_{i-1})$, thereby reducing the word history to the preceeding $n-1$ words. A back-off mechanism is generally used to smooth the estimates of the probabilities of rare *n*-grams by relying on a lower order *n*-gram when there is insufficient training data, and to provide a means of modeling unobserved word sequences [15]. While 3- and 4-gram LMs are the most widely used, class-based *n*-grams, and adapted LMs are recent research areas aimed at improving LM accuracy.

Acoustic feature extraction is concerned with the choice and optimization of acoustic features in order to reduce model complexity while trying to maintain the linguistic information relevant for speech recognition. Acoustic modeling must take into account different sources of variability present in the speech signal: those arising from the linguistic context and those associated with the non-linguistic context such as the speaker and the acoustic environment and recording channel. Most state-of-the-art systems make use of hidden Markov models (HMMs) for acoustic modeling, which consists of modeling the probability density function of a sequence of acoustic feature vectors. The most widely used solutions model context-dependent phones and use a host of techniques such as parameter sharing, feature analysis, linear and non-linear transformation, noise compensation and discriminative training to improve model accuracy. Regarding the training data, the first 100-200 hours of representative data provide the most gain for acoustic modeling, with additional data giving only small improvements.

The pronunciation lexicon is the link between the representation at the acoustic-level (frames of features) and at the word level. At the lexical and pronunciation level, two main sources of variability are the dialect and individual preferences of the speaker. There are three main steps in designing a recognition lexicon: definition and selection of the vocabulary items, representation of each pronunciation entry using the basic acoustic units of the recognizer, and estimation of probabilities for pronunciation variants. Lexical coverage has a large impact on recognition performance, and the accuracy of the acoustic models is linked to the consistency of the pronunciations in the lexicon. The recognition vocabulary is usually selected to maximize lexical coverage for a given size lexicon. Since on average, each out-of-vocabulary (OOV) word causes more than a single error (usually between 1.5 and 2 errors), word list selection is an important design step. At LIMSI, word list selection is carried out by choosing the n most probable words after linear interpolation of unigram LMs trained on the different text sources so as to maximize the coverage on a set of development data. The vocabulary size, n is chosen so as to minimize the OOV rate while keeping a reasonable size and avoiding typos. The lexicon typically contains canonical pronunciations and frequent variants, which are generated either manually or by rule. Sometimes non-speech events and compound words or short phrases are also explicitly included as lexical entries.

Given the speech signal and the models (lexicon, acoustic and language), the job of the decoder is to determine the word sequence with the highest likelihood (MAP decoding) or maximizing the expected accuracy of the hypothesis (consensus decoding). The main decoding challenge for large vocabulary continuous speech recognition (LVCSR) is to design an efficient algorithm to explore the huge search space, for which it is generally impossible to carry out an exhaustive search. Many techniques have been proposed to reduce the needed computation by limiting the search space [6]. It has become common practice to use multi-pass decoding strategies which can limit the complexity of each individual decoding pass, allowing more complex models (additional knowledge) to be used progressively. Information is usually transmitted between passes via word graphs, containing the word hypotheses and their respective scores.

Table 1. Indicative speech recognition word error rates for different tasks and speaking styles

Task	*Condition*	*Word Error*
Dictation	read speech, close-talking mic.	3-4% (humans 1%)
	read speech, noisy (SNR 15dB)	10%
	read speech, telephone	20%
	spontaneous dictation	14%
	read speech, non-native	20%
Found audio	TV & radio news broadcasts	10-15% (humans 4%)
	documentaries	20-30%
	European Parliament	8%
	telephone conversations	20-30% (humans 4%)
	lectures (close mic)	20%
	lectures (distant mic)	50%

Table 2 gives some indicative word error rates for a range of speech recognition tasks and speaking styles. For a few of the tasks some measures of human performance are available. Studies comparing human and machine transcription performance [26,11,22] show that humans consistently do considerably (5 to 10 times) better than machines.

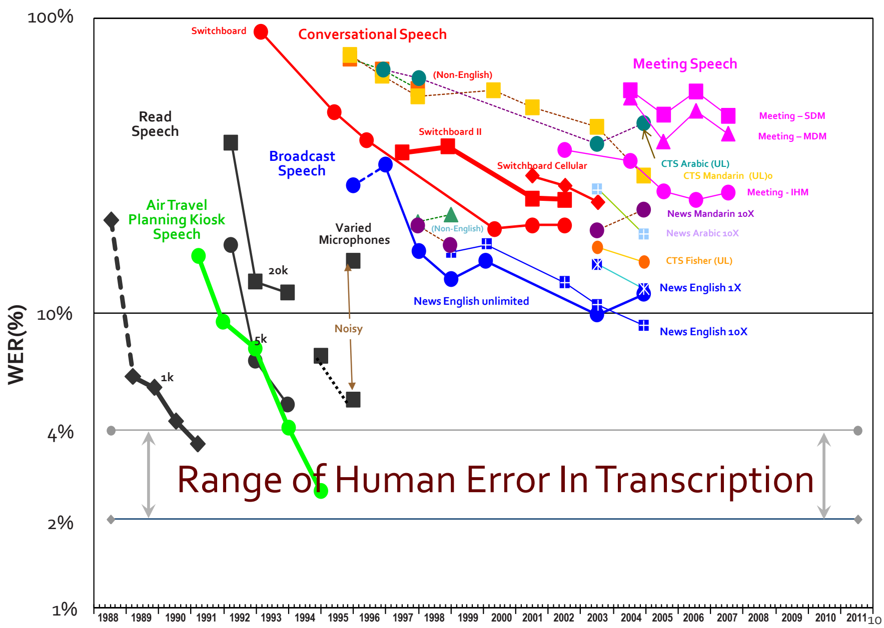

Fig. 2. NIST summary of automatic speech recognition evaluation history (May'07). The word error rate (WER) of the best system for each evaluation/task is shown. (Figure reprinted from [24].

The results in the top part of the table are for a dictation task, where under ideal conditions (i.e. the text is already prepared and the speaker uses a close-talking microphone in a quiet acoustic environment, and the goal is to speak to the machine), quite low error rates can be achieved. It can be noted however than even in this situation human performance is much better than the machine. Any perturbation, such as a noisy environment, a telephone acoustic channel, or accented speech from a non-native speaker results in a very significant increase in error rate. There is also a much higher error rate if the speaker does not read a text, but rather prepares the subject and formulates the ideas on the fly as shown by the entry labeled spontaneous dictation. The lower part of the table reports performance on some different types of 'found data,' that is data that was produced for independent purposes, but have been of interest to the research community since there are a range of potential applications that can be enabled via speech processing technologies. Broadcast data has been attracting growing interest since the task was introduced by DARPA over a decade ago. While initial error rates were quite high, today word error rates in the range of 10-15% have been reported on broadcast news data for a number of languages (English, French, Spanish, German, Dutch, Arabic, Mandarin, Portuguese, Japanese). A wide range in performance is observed for different data types, with quite low error rates for the speech of main announcers in recording studios, and much higher error rates for distant reporters, particularly when the acoustic channel or environment is poor. Similarly, the error rate can increase dramatically if the interactivity is high (interviews, debates). Documentaries are particularly challenging to transcribe, as the audio quality is often not very high, and there is a large proportion of voice over. The recent TC-STAR project (`http://www.tc-star.org`), which targeted speech-to-speech translation of unconstrained conversational speech from the European Parliament Plenary Sessions (EPPS), reported word error rates of about 8% for European English and Spanish. Word error rates on conversational telephone speech and lectures (`http://chil.server.de`) and meetings (`http://www.amiproject.org`) are substantially higher, reflecting some of the additional challenges of these domains.

Figure 2 summarizes the results of NIST sponsored benchmark speech recognition evaluations over the last 20 years. Each curve corresponds to a specific task, and plots the word error rate of the best system in each evaluation. The first evaluations were for read speech, with a move in the mid 1990's to conversation telephone speech and to broadcast data. Over time the amount of data used to train the speech recognizers increased along with model complexity (and vocabulary size). It can be seen that typically as the performance of the best systems approached 10%, more challenging tasks were introduced. The performance of humans is significantly better than that of machines for all types of real-world data.

3 Enriched Transcription

The speech signal encodes both the linguistic message and other types of information such as the characteristics of the speaker, the acoustic environment, the recording conditions and the transmission channel. Ideally we would like to identify as many of these characteristics as possible from the audio channel. For example, a first processing step can partition the audio signal, extracting acoustic-based meta-data and creating a description of the audio document in terms of the language(s) spoken, the speaker(s), accent(s), acoustic background, speaker's emotional state etc. Such information can be used to improve speech recognition performance, and to provide an enriched text output for downstream processing. The automatic transcription can also be used to provide information about the linguistic content of the data (topic, named entities, speech style, ...). By associating each word and sentence with a specific audio segment, an automatic transcription can allow access to any arbitrary portion of an audio document. If combined with other meta-data (language, speaker, entities, topics) access via other attributes can be facilitated. Enriched transcription also includes the inclusion of case and punctuation in the output.

Language and speaker recognition make use of similar modeling techniques as those use for speech recognition. There are two predominant approaches to language recognition, acoustic (Gaussian mixture) models) and phonotactic models [33]. Both types of systems require only untranscribed training data each target language of interest, but phonotactic-based systems are somewhat less sensitive to changes in recording conditions. Other techniques such as Support Vector Machines and system combination (fusion) have also been proposed. Speaker recognition [10] is the process of identifying a speaker from their voice. Two tasks are typically distinguished, speaker identification and speaker verification. For the former, the speaker is identified as one of a set of known speakers (closed task) or as none of them (open task). For the second task, given a speech sample, the system needs to decide if the sample was produced by a given speaker (the decision is yes or no). NIST (http://www.nist.gov/speech/tests) has been organizing language and speaker recognition benchmarks for conversational telephone speech since 1996.

Speech-to-text systems historically produce a case insensitive, unpunctuated output. In the context of the TC-STAR project tools to automatically add case and punctuation were developed [18]. Both linguistic and acoustic information (essentially pause and breath noise cues) are used to add punctuation marks in the speech recognizer output. This is done by rescoring a word lattice that has been expanded to permit punctuation marks after each word, sentences boundaries at each pause, with a specialized case sensitive, punctuated language model.

Speaker diarization, also referred to speaker segmentation and clustering, has been of recent interest to the speech community. It is a useful preprocessing step for an automatic speech transcription system, in that it enables unsupervised speaker adaptation to be carried out at a cluster level, thus increasing the amount of available data which can improve transcription performance. The performance of diarization systems has been assessed in the Rich Transcription benchmarks(http://www.nist.gov/speech/tests/rt) under the DARPA EARS program, as well as in the CHIL, AMI and ESTER evaluation campaigns. One of the major issues is that the number of speakers is unknown a priori and needs to be automatically determined. In [8,31] speaker recognition techniques were shown to improve the performance of a diarization system. In these evaluations the goal was to correctly attribute speech segments to unidentified speakers in the audio document, that is there was no attempt to determine the true identity of the speaker.

Speaker diarization can also improve the readability of an automatic transcription by structuring the audio stream into speaker turns, in some cases by providing the true speaker identity. For example, in broadcast news programs, the speaker names are often explicitly stated, providing the true identities of those taking part in the show. A future aim is to combine speaker recognition techniques to identity speakers from a very large population. One of the goals in the QUAERO project (http://www.quaero.org) is to explore the novel use of the linguistic information produced by a speech recognizer to complement the information derived from the acoustics. The main idea of the 'Who's Who' procedure is to exploit the structure of broadcast data to automatically learn the names of speakers in a large unannotated corpus without the need for human intervention.

4 Some Recent Progress and Outstanding Challenges

One of the challenges for automatic language processing is the portability of technology across languages. Multilinguality is of particular interest for Internet-based applications, where information may first (or only) be available in another language than the user's mother tongue. A recent book [2] addresses issues in multilingual speech processing. Word error rates below 20% were reported for a number of languages [21]. With appropriately trained models, recognizer performance was observed to be more dependent upon the type and source of data, than on the language.

Speech recognizers for well-covered languages are typically trained on hundreds of hours of transcribed speech and hundreds of millions of words of texts. Thus data collection and preparation require significant investment, in terms of money, time and human effort. Reducing these costs is an important research direction (http://coretex.itc.it). For acoustic modeling, it has been proposed to use a speech recognizer [16,20,30] to reduce transcription costs. For some applications iterative training using automatic transcripts may be sufficient, whereas in other cases a human may need to correct the transcription. In the context of the DARPA EARS (http://w2.eff.org/Privacy/TIA/ears.php) program, extensive experiments were reported using 'quick' transcriptions to reduce the human annotator time for a conversational telephone speech task [17]. The approach has also adopted for use in the DARPA GALE (http://www.darpa.mil/ipto/programs/gale/gale.asp) program in order to reduce transcription costs and therefore provide more data. Acoustic model training requires an alignment between the audio signal and the phone models, which usually relies on a perfect orthographic transcription of the speech data and a good phonetic lexicon. Making use of these quick transcriptions has led to revisions in acoustic model training procedures to make them more flexible [25] and less dependent on a perfect transcription.

Obtaining resources is particularly difficult for 'lesser' represented languages that do not have a strong strategic (economic or security) push. Language preservation is important for cultural diversity, and transmission of cultural heritage (http://cmuspice.org, http://projects.ldc.upenn.edu/LCTL). A recent workshop addressed the topic of developing spoken languages technologies for under-resourced languages [1]. Given recent trends for computerization, such languages pose many new research challenges. In general it is relatively easy to obtain audio data, by recording radio or television programs. Finding text material in electronic form is often more difficult since many languages are poorly represented, if at all, on the Internet. For some languages there are no commonly adopted writing conventions or there may have been recent writing reforms which result in quite varied text materials. Another complication is that it is difficult to find people that have expertise in both the language of interest and in language processing. Written resources and a pronunciation dictionary are the most critical for todays technologies: reasonable acoustic models can be trained on several tens to hundreds of hours of data which can be obtained at a reasonable cost. Given that economic or political reasons are unlikely to support the development of technologies for many of these lesser languages, likely viable solutions will rely on new lightly supervised or unsupervised training techniques. Some work in this direction has been reported in [23,5] for pronunciation modeling, and a framework for the development of resources and models is being developed in the SPICE project (http://cmuspice.org). As mentioned earlier, only about 15% of the world's languages are written, so current word based modeling techniques cannot be directly applied to the remaining languages. For relatively small data collections, approaches based on phone-like units may provide a short-term solution for such languages [29].

Concerning language modeling, as the amount of available data has increased, most state-of-the-art systems use back-off n-gram language models which result from the interpolation of language models trained on non-overlapping subsets of the available language model training material. This allows different interpolated weights to be associated with different data subsets, thus increasing or reducing their importance. The interpolation weights are optimized on a set of development data. It is often the case that the

Table 2. Observed pronunciations for four inflected forms of the word 'interest' in American English broadcast news (BN) and conversational telephone speech (CTS) data

Word	*Pronunciation*	*BN*	*CTS*	*Word*	*Pronunciation*	*BN*	*CTS*
interest	In**trI**st	238	488	interests	In**trI**ss	52	53
	In**tXI**st	3	33		In**trI**sts	19	30
	In**XI**st	0	11		In**tXI**sts	3	2
					In**tXI**ss	3	1
interested	In**trI**stxd	126	386	interesting	In**trI**st|G	193	1399
	In**tXI**stxd	3	80		In**tXI**st|G	8	314
	In**XI**stxd	18	146		In**XI**st|G	21	463

vast majority of training texts come from written sources (newspapers, newswires, ...), and audio transcripts represent only a small portion of the data. In the LIMSI Arabic speech-to-text system, the coefficients associated with the audio transcriptions, account for almost 0.5, even though these texts represent only about 1% of the available data. This highlights the importance of audio transcripts for language modeling of speech.

Although proposed a decade ago[13], Multi-Layer Perceptron (MLP) features have recently been attracting interest for large vocabulary speech recognition due to their complementarity with cepstral features [32]. Even though probabilistic features have never been shown to consistently outperform cepstral features in LVCSR, having different properties they can markedly improve the performance when used in conjunction with them. Connectionist models have also been shown to be effective for language modeling [28].

Concerning pronunciation modeling, most of todays state-of-the-art systems include pronunciation variants in the dictionary, associating pronunciation probabilities with the variants [3,4]. However, for large vocabulary systems most of the lexical items are never or only rarely observed. Table 4 shows the observed pronunciation counts for four inflected forms of the word 'interest' in about 100 hours American English broadcast news and conversational telephone speech data. It can be seen that the number of occurrences varies quite a bit for the different forms, and the data type. As can be expected there is a higher proportion of reduced forms are observed in CTS data than in BN data. Two main reductions are observed: the transformation of 'ter' into 'tr' (loss of the schwa) and the deletion of the 't' ('inter' is realized as 'iner'). In the recognition dictionary there are a number of similar, less frequent words: interestingly, disinterest, disinterested for which it would be nice to predict pronunciation variants, as well as for other words with a similar syllabic structure: interfere, interfering, interconnect, intercom, ... So an unresolved problem is how to accurately model pronunciation variants. It has been observed that a person will pretty much systematically choose a pronunciation variant, so one research direction is to develop style-specific or accent-specific pronunciations models, which could be adapted to a particular speaker.

Unsupervised model adaptation has been demonstrated to be quite successful for acoustic modeling, and is widely used in most state-of-the-art transcription systems. Several directions have been explored for adaptive language modeling with less convincing results [27]. Concerning pronunciation modeling, large amounts of data are needed to estimate accurate pronunciation probabilities. Where for acoustic modeling a few minutes of speech provides a fair amount of acoustic data for adaptation, this data only contains a few hundred words, many of which do not carry much information content. There are a few more phones for pronunciation modeling, however most are unlikely to be distinctive of the speaker/dialect.

5 Conclusion

Automatic speech recognition is a key technology for audio indexing. Recent progress has enabled the development of systems for a handful of languages that achieve word errors rates the order of 10 to 30% depending upon the type of data. Such performance levels are sufficient to support some near-term applications for structuring and mining spoken data collections, in particular those containing prepared speech. Higher error rates on the order of 20-50% have been reported for speech data from more interactive situations (interviews, debates, conversations, meetings). Transcriptions of speech data remain critical for language modeling, since 100 hours represents only about 1 million words of texts which is largely insufficient. Some recent efforts have been devoted to locating speech-like texts on the Internet.

References

1. International Workshop on Spoken Languages Technologies for Under-resourced languages, SLTU Hanoi, (May 2008), `http://www.mica.edu.vn/sltu`
2. Schultz, T., Kirchhoff, K. (eds.): Multilingual Speech Processing. Elsevier, Amsterdam (2006)
3. Bourlard, H., Furui, S., Morgan, N., Strik, H. (eds.): Modeling pronunciation variation for automatic speech recognition.In: Speech Communication, vol. 29(2-4) (November 1999) (Special issue)
4. Fosler-Lussier, E., Byrne, W., Jurafsky, D. (eds.): Pronunciation Modeling and Lexicon Adaptation.In: Speech communication, vol. 46(2) (June 2005) (Special issue)
5. Adda-Decker, M., Lamel, L.: Pronunciation variants across system configuration, language and speaking style. Speech Communication 29(2-4), 83–98 (1999)
6. Aubert, X.L.: An overview of decoding techniques for large vocabulary continuous speech recognition. Computer Speech & Language 16(1), 89–114 (2002)
7. Bahl, L.R., Baker, J.K., Cohen, P.S., Dixon, N.R., Jelinek, F., Mercer, R.L., Silverman, H.F.: Preliminary results on the performance of a system for the automatic recognition of continuous speech. In: IEEE ICASSP-1976, Philadelphia (April 1976)
8. Barras, C., Zhu, X., Meignier, S., Gauvain, J.L.: Multistage speaker diarization of broadcast news. IEEE Transactions on Audio, Speech and Language Processing 14(5), 1505–1512 (2006)
9. Bulyko, I., Ostendorf, M., Stolcke, A.: Gtting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In: Hearst, M., Ostendorf, M. (eds.) HLT-NAACL 2003, Edmonton, March 2003, vol. 2, pp. 7–9 (2003)
10. Campbell, J.: Speaker Recognition: A Tutorial. Proc. of the IEEE 85(9) (September 1997)
11. Deshmukh, N., Duncan, R., Ganapathiraju, A., Picone, J.: Benchmarking Human Performance for Continuous Speech Recognition. In: Fourth International Conference on Spoken Language Processing, Philadelphia, October 1996, vol. 1(10) (1996)
12. Gauvain, J.L., Lamel, L., Adda, G.: The LIMSI Broadcast News Transcription System. Speech Communication 37(1-2), 89–108 (2002)

13. Hermansky, H., Sharma, S.: TRAPs - classifiers of TempoRAl Patterns. In: ICSLP 1998, Sydney (November 1998)
14. Jelinek, F.: Continuous Speech Recognition by Statistical Methods. Proc. of the IEEE 64(4), 532–556 (1976)
15. Katz, S.M.: Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. IEEE Trans. Acoustics, Speech & Signal Processing ASSP-35(3), 400–401 (1987)
16. Kemp, T., Waibel, A.: Unsupervised Training of a Speech Recognizer: Recent Experiments. In: ESCA Eurospeech 1999, Budapest, Hungary, September 1999, vol. 6, pp. 2725–2728 (1999)
17. Kimball, O., Kao, C.L., Iyer, R., Arvizo, T., Makhoul, J.: Using Quick Transcriptions to Improve Conversational Speech Models. In: ICSLP 2004, Jeju, (October 2004)
18. Lamel, L., Gauvain, J.L., Adda, G., Barras, C., Bilinski, E., Galibert, O., Pujol, A., Schwenk, H., Zhu, X.: The LIMSI 2006 TC-STAR EPPS Transcription Systems. In: ICASSP, Honolulu, April 2007, pp. 997–1000 (2007)
19. Lamel, L., Gauvain, J.L.: Speech Recognition. In: Mitkov, R. (ed.) Chapter 16 in *OUP Handbook on Computational Linguistics*, pp. 305–322. Oxford University Press, Oxford (2003)
20. Lamel, L., Gauvain, J.L., Adda, G.: Lightly Supervised and Unsupervised Acoustic Model Training. Computer, Speech & Language 16(1), 115–229 (2002)
21. Lamel, L., Gauvain, J.L., Adda, G., Adda-Decker, M., Canseco, L., Chen, L., Galibert, O., Messaoudi, A., Schwenk, H.: Speech Transcription in Multiple Languages. In: IEEE ICASSP 2004, Montreal (April 2004)
22. Lippmann, R.P.: Speech recognition by machines and humans. Speech Communication 22(1), 1–16
23. Pellegrini, T., Lamel, L.: Experimental detection of vowel pronunciation variants in Amharic. In: LREC 2006, Genoa (2006)
24. Przybocki, M.: Technology Advancements have Required NIST Evaluations to Change Data and Tasks - and now Metrics. In: Presented at the ELRA Workshop on Evaluation, LREC 2008, Marrakesh (2008)
25. Stolcke, A., Chen, B., et al.: Recent innovations in speech-to-text transcription at SRI-ICSI-UW. IEEE Transactions on Audio, Speech, and Language Processing 14(5), 1729–1744 (2006)
26. van Leeuwen, D.A., van den Berg, L.G., Steeneken, H.J.M.: Human Benchmarks for Speaker Independent Large Vocabulary Recognition Performance. In: ESCA Eurospeech 1995, Madrid, pp. 1461–1464 (September 1995)
27. Rosenfeld, R.: Two decades of statistical language modeling: where do we go from here? Proc. IEEE 88(8), 1270–1278 (1999)
28. Schwenk, H.: Continuous space language models. Computer Speech and Language 21, 492–518 (2007)
29. Van Thong, J.M., Goddeau, D., Litvinova, A., Logan, B., Moreno, P., Swain, M.: SpeechBot: a speech recognition based audio indexing system for the web. In: RIAO 2000 Content-Based Multimedia Information Access, Paris, pp. 106–115 (April 2000)
30. Zavaliagkos, G., Colthurst, T.: Utilizing Untranscribed Training Data to Improve Performance. In: DARPA Broadcast News Transcription & Understanding Wshop (November 1998)
31. Zhu, X., Barras, C., Lamel, L., Gauvain, J.L.: Speaker Diarization: from Broadcast News to Lectures. In: Renals, S., Bengio, S., Fiscus, J. (eds.) MLMI 2006. LNCS, vol. 4299, pp. 396–406. Springer, Heidelberg (2006)
32. Zhu, Q., Stolcke, A., Chen, B.Y., Morgan, N.: Using MLP features in SRI's conversational speech recognition system. Interspeech 2005, 2141-2144, Lisbon (2005)
33. Zissman, M.A.: Comparison of Four Approaches to Automatic Language Identification of Telephone Speech. IEEE Trans. Speech and Audio Proc. 4(1), 31–44 (1996)