

HUMAN LANGUAGE TECHNOLOGY: OPPORTUNITIES AND CHALLENGES

*Mari Ostendorf*¹ *Elizabeth Shriberg*^{2,3} *Andreas Stolcke*^{2,3}

¹University of Washington ²SRI International ³International Computer Science Institute
mo@ee.washington.edu {ees, stolcke}@speech.sri.com

ABSTRACT

In recent years, there has been dramatic progress in both speech and language processing, in many cases leveraging some of the same underlying methods. This progress and the growing technical ties motivate efforts to combine speech and language technologies in spoken document processing applications. This paper outlines some of the issues involved, as well as the opportunities, presenting an overview of the special double session on this topic.

1. INTRODUCTION

Human language technology (HLT) provides important tools for making use of the vast amount of information in documents available via the web, and significant recent progress has been made in areas such as text retrieval, analysis, summarization and translation. While much of this work has focused on text documents, speech and video signals are also increasingly available. We refer to such signals – including TV and radio broadcasts, congressional records, oral histories, voicemail, call center recordings, etc. – as “spoken documents”. As speech recognition technology improves, language processing for spoken audio has attracted increased interest. And because it takes longer to listen to audio than to read text, spoken documents are clearly a prime candidate for automatic indexing, information extraction, and other such technologies.

Over the last decade, the speech processing and natural language processing communities have developed largely independently, though many of the algorithms stem from the same fundamental theory. With the growing importance of spoken document processing, there is now a need to bridge this gap. This session takes a step towards this goal, by introducing speech researchers to downstream applications that could be applied to speech (and video), and by providing language processing researchers with insights into what speech has to offer beyond word information. Many of the papers in this session raise issues in applying text-based technologies to spoken documents. The differences between written and spoken documents have implications for both speech and language processing modules. In addition, since HLT is ultimately needed for human processing of information, we include two papers on assessing of the impact of technology on human performance in various information processing tasks.

Despite little interaction between the speech and language processing communities, there has been some technology exchange through work in dialog systems, and both communities are leveraging advances in machine learning. Hence, we expect the session will also bring to light a wealth of shared algorithmic methods that could be useful in both fields, and where cross-fertilization is likely to provide mutual benefits. A few such shared techniques are highlighted in this overview; we encourage our readers to search for more examples in the papers in this session.

The goal of this paper is to set the context for the session, providing background on the various technologies and raising issues that cut across the relevant fields. In Section 2 we give an overview of the state-of-the-art in large vocabulary speech recognition, to provide perspective on what might be available for spoken document processing. In Sections 3 and 4, we outline issues in speech processing that impact language processing, including information beyond the words and methods for handling speech recognition errors. In Section 5 we discuss some common threads in the methods used in speech and language processing. Finally, in Section 6 we provide an overview of the eleven invited papers in this special double session.

2. LARGE VOCABULARY SPEECH RECOGNITION

Most HLT applications require the ability to accurately transcribe unrestricted, open-vocabulary speech. Since the mid-1990s, progress in large vocabulary recognition has been driven by annual evaluations conducted by NIST for automatic transcription of broadcast news (BN), conversational telephone speech (CTS), and recently multi-party meetings [1]. Evaluation conditions have become more difficult over the years, by the imposition of factors such as runtime limits, automatic segmentation requirements, and broadening of data sources. Nevertheless, word error rates (WERs) have declined from around 30% for BN and above 50% for CTS, to below 10% and 15%, respectively. These improvements are due in part to availability of increasing amounts of training data, which now comprise more than 2000 hours for both English BN and CTS. But there have been many research achievements as well, including techniques that make use of cheaper and therefore larger data sources (e.g. training on errorful transcriptions). In addition, the availability of more data has spawned the development of more sophisticated models. The systems have achieved remarkable convergence, across both sites and domains. In the paragraphs below, we overview key elements typically found in the NIST-evaluated systems.

Front ends use cepstral analysis in combination with dimensionality reduction techniques, such as heteroscedastic LDA, starting from up to third-order delta features or the concatenated cepstral vectors from several adjacent frames. Recent developments are discriminatively trained feature extraction methods such as fmPE [2] or multi-layer perceptrons [3]. A host of techniques are used to reduce mismatch between trained models and test data, and to reduce inter-speaker variability in training. Standard techniques include vocal-tract length normalization, adaptation of acoustic models using maximum likelihood linear regression (MLLR), and speaker-adaptive training based on MLLR. The acoustic models are mixtures of Gaussians, typically with several hundred thousand to a million distributions with diagonal covariances; recently

systems have started to use full covariances and fewer Gaussians. Distributions are clustered by decision trees, using phone context and other features. Once clustered, Gaussians are trained using discriminative criteria such as maximum mutual information or minimum phone error [4], which reduce WER over maximum likelihood training.

Language modeling is dominated by four- and five-gram models, typically kept to manageable size by entropy-based pruning. Recent developments with grammar-based language models have not found widespread use yet, mainly due to computational constraints. A new development is trainable, continuous vector space representations for the vocabulary that employ neural networks as smooth conditional probability estimators [5]. Attempts to adapt the LM to topics, dialog structure, or other higher-level aspects of the domains have shown only very marginal improvements so far, and are not generally used. Instead, one typically builds separate LMs from a variety of sources, including web text collections [6], and interpolates them with globally optimized weights.

Decoding typically proceeds in multiple stages, allowing progressively more expensive models, and iterative normalization and adaptation. Cross-adaptation makes use of multiple recognizers that differ in signal processing, pronunciation models or acoustic modeling approaches, and lets one subsystem adapt to the output of another. This helps avoid the reinforcement of recognition errors, and is similar to co-training used for weakly supervised training of taggers and parsers [7]. Finally, the outputs of different subsystems and decoding stages are combined by tallying up “votes” (i.e., posterior probabilities) for competing word hypotheses.

Despite significant advances, it is still true that mismatches between training and test data in terms of acoustics or speaking style degrade recognition accuracy considerably, for example, when recognizing new genres of broadcast shows, or when running a CTS-based recognizer on meetings. Better robustness and portability thus remain key goals for future research. In addition, performance on non-English speech is typically a couple of years behind that in the corresponding English tasks, due to the longer history of work on English, greater availability of resources, and differences in the languages themselves.

3. SPEECH VS. TEXT: BEYOND WORDS

Spoken language differs from text in terms of stylistic factors [8], as well as in terms of what information is conveyed explicitly. Most notably, spoken language does not contain explicit punctuation, capitalization, or formatting. If the spoken document involves spontaneous speech (e.g. meetings, conversations), then high rates of disfluencies (e.g. filled pauses, restarts, repetitions and self-corrections) are often present [9]. Such factors pose difficulties for automatically processing spoken documents, though humans do not find it particularly difficult. Of course, humans can make use of semantics, context and world knowledge that is far beyond what current systems can model. However, humans also make use of acoustic cues to speaker identity, emphasis and structural organization of the words. Such information is not typically expressed in the transcripts produced by a speech recognition system, but it may be easier to extract and represent than semantic and pragmatic knowledge. Hence, many researchers believe that automatic language processing could benefit from a richer representation of the audio signal that incorporates this information. Efforts at defining such a representation, promulgated as “Rich Transcription” by the DARPA EARS program, are in their early stages, but evidence for

its utility is growing.

Several types of beyond-words information, or “metadata”, can enrich the representation of speech. For example, audio diarization, annotation of structural information (analogous to punctuation in written text), and annotation of higher level discourse information such as dialog acts and topic boundaries. In the EARS program, the term “metadata” has to date referred to information that can be construed (more or less) by a human listener who only has access to the audio. We note that the term “metadata” is also used elsewhere in the speech and language community to include supplemental information, such as radio production notes [10] or thesaurus terms [11]. In this case, the information is not encoded in a person’s voice, but it is also of use for downstream language processing applications. In this paper, we focus on information available from the speech signal, because of the connections to and reliance on signal processing.

Audio diarization critically includes indexing of speakers [12], since speaker information is important for resolution of some pronouns and provides a cue to topic and sentence boundaries. In some applications it is also of interest to label particular speakers, for example key political figures, since interpretation and use of spoken information can depend on who the source is. Other types of audio information that may be of interest include channel changes (e.g., telephone channel for a call-in speaker), music, advertising segments, and so on.

For lower-level structural metadata, one could use a surface-form representation that imitates orthographic transcription, e.g., automatically detected punctuation plus hyphens to mark self-correction points [13]. Alternatively, one could represent the underlying structure (interruption points, edit regions, boundaries of sentence-like units (SUs)) [14], which is closer to the acoustic cues in the signal and richer in terms of the representation of disfluencies, but requires some revision of text processing technology to handle the new representation. We argue that speech is sufficiently different from text that there are modifications needed in models in any case, and hence there may be only a small added cost to using the richer structural representation. Further, many of the techniques required for detecting surface form punctuation involve language processing technology that is arguably better left to the end application. That said, many results demonstrate that some use of language cues is critical to achieving good performance in detection of structural metadata [14, 15, 16].

Higher-level metadata, such as labeling of topics, dialog acts, emotion, speaker state, and so on, is not currently a focus in the DARPA “Rich Transcription” paradigm. But in the broader community, there is significant and growing interest in recovering such information from speech. Automatic detection of dialog acts is important for intelligent human-computer interfaces [17] and for understanding human-human conversations [18, 19]. Both topic boundaries [20] and dialog acts [21, 18, 19] are a focus in recent work on meeting understanding. Emotion detection [22], once a “fringe” topic in speech processing, is now of great interest for such varied applications as call-center triage, games, tutoring systems, in-car navigation systems, and even home health monitoring.

For all of these tasks, prosodic features (duration, fundamental frequency, energy and voice source characteristics) can play a helpful role. Prosodic cues mark various levels of segmentation and salience in speech, as well as intent (as in distinguishing questions vs. statements). Prosodic cues have also proved to be useful for speaker identification [23]. A challenge in modeling prosody, because it plays a role at many levels, is factoring out or normaliz-

ing features to account for effects at different levels.

Since text processing systems productively make use of punctuation, it makes sense to use detected structural metadata in automatic language processing systems operating on speech, and there are a few examples indicating that such information is useful. Work in parsing spontaneous speech has shown that detection of interruption points and edit regions is needed for good performance, e.g., [24, 25], but incorporating (quantized) sentence-internal prosody cues as words is not as effective as using punctuation [26]. However, this work assumes known sentence boundaries. When sentences must be automatically segmented, there is a clear benefit for parsing from using explicit metadata detection algorithms rather than a simple pause-based segmentation [27]. Work in dialog systems has successfully incorporated prosodic metadata for both parsing and control of the dialog manager [28].

4. HANDLING ERRORFUL TRANSCRIPTS

On top of missing orthographic cues and stylistic differences, the transcripts produced by a speech recognition system have errors (as surveyed above) and often important names and places are missed because they are absent from the recognizer vocabulary. Speech and language researchers working on dialog systems have also long recognized that better understanding results can be obtained if more than one hypothesis is passed from the recognizer to the understanding module. Options for representing recognizer alternatives include N-best sentence hypotheses, word lattices, confusion networks [29], or simply augmenting a single hypothesis with word confidence estimates.

None of these alternatives addresses the problem of out-of-vocabulary (OOV) words, since the correct word will not appear in any of the alternative hypotheses. The paper in this session by Van Thong *et al.* [10] considers possible representations for handling this problem within the context of a retrieval application. Other results in named entity detection [30] and dialog systems [31] indicate that language context combined with recognizer confidence estimates can provide sufficient clues to detect names, even when these are obscured by recognition errors. Once a candidate OOV word is detected, the system can back off to a phonetic representation to allow for new vocabulary items [32, 33].

5. COMMON THEMES

Many of today's HLT systems use a statistical approach, typically involving (among other things) a language model to estimate the probability of a given sequence of words. In speech recognition, a language model is used to distinguishing between acoustically confusable word sequences and to bias the result to be consistent with the target task domain. In machine translation and text summarization, a language model is used to rank the outputs of a translation (or compression) model and generate well-formed sentences. In information extraction, a language model may be used to characterize word sequence cues to names and other important information. Recent work in video annotation has explored cross-media language models [34]. In most systems, the n-gram initially explored in speech recognition is the model of choice because of its relative simplicity, but in some applications (including metadata extraction [14]) there is increasing interest in alternative approaches, e.g. conditional random fields and maximum entropy models.

Consistent with increased use of statistical techniques, language processing technology has moved with speech technology towards reliance on data-driven learning with ever-increasing demands for data. As demand outstrips the availability of task-dependent annotated data, researchers have turned towards weakly supervised and unsupervised training techniques – a highlighted theme in the past two HLT-NAACL meetings. In addition, researchers have looked towards methods of harvesting text data from the web and other easily available sources [6, 35, 36].

Speech and language research also have in common the emphasis on evaluation metrics. Speech recognition has long been driven by the goal of minimizing word error rate, and the parsing community has adopted a standard metric for several years now, which arguably led to the significant advances in that field. However, the subjectivity of language has posed problems in defining standard scoring metrics for tasks such as translation, generation, and summarization. With the introduction of the BLEU score [37] in machine translation, there is now much interest in quantitative evaluation metrics for this and other HLT areas. Importantly, the availability of quantitative scoring metrics accelerates the research process and makes it possible to automatically optimize methods for combining knowledge sources. The idea of automatically learning strategies for knowledge combination has long used in speech recognition, but is now being applied in other HLT tasks, e.g., parsing [38, 15] and machine translation [39].

There are many other important areas where parallels exist, including dimensionality reduction techniques and fast search algorithms. In addition, there are text processing technologies that could potentially benefit speech recognition, such as transliteration for text normalization in language model training and morphological analysis for language model decomposition. Space limits prevent an extensive survey, and such a survey would be quickly outdated. Instead, we provide these examples to motivate the need for more extensive cross-fertilization between disciplines.

6. SESSION OVERVIEW

The papers in this double session address core technology in both speech and language processing, as well as applications and impact on human users. The first two papers address the problem of richer annotation of the word stream in speech processing, including both audio diarization [12] and detection of structural metadata [14]. The third paper [15] provides a bridge between speech and language technologies, reviewing statistical parsing and discussing its role in both speech recognition and in detection of disfluencies. The next three papers review language processing technology in three key areas: translation [39], information extraction [40], and summarization [41]. These technologies are important for information management and have a long history of text-based processing. The papers present a review of the basic technology, as well as raise issues in moving to spoken documents.

Next are two contributions that look more directly at integrating speech and language in indexing and retrieval of audio [10] and video [34] signals. These papers present an overview of current technology and describe both challenges and opportunities associated with dealing with large archives of audio and video documents. Following are two papers that put humans (the ultimate consumer of most language technology) in the loop. One [42] investigates the impact of recognition errors and metadata detection on human readability and comprehension, using both speech transcripts and machine translation output. The other [43] puts for-

ward a new evaluation paradigm for machine translation that incorporates human users in a realistic document retrieval task. The final paper of the session [44] deals with human-computer dialog systems, but focuses on the goal of portability to new tasks and domains. Many of the issues raised are universal to a broad range of HLT applications in both speech and language processing.

Acknowledgments

This research has been supported by DARPA under contract MDA972-02-C-0038, and NSF under contracts IIS-0085940 and IIS-032676. Distribution is unlimited. Any opinions expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA or NSF.

7. REFERENCES

- [1] National Institute of Standards and Technology, “RT-03F evaluation,” <http://www.nist.gov/speech/tests/rt/rt2003/fall/rt03f-evaldisdoc/index.htm>, 2003.
- [2] D. Povey et al., “fmPE: Discriminatively trained features for speech recognition,” in *Proc. DARPA Rich Transcription Workshop*, 2004.
- [3] Q. Zhu et al., “Incorporating Tandem/HATs MLP features into SRI’s conversational speech recognition system,” in *Proc. DARPA Rich Transcription Workshop*, 2004.
- [4] D. Povey and P. C. Woodland, “Minimum phone error and I-smoothing for improved discriminative training,” in *Proc. ICASSP*, 2004.
- [5] H. Schwenk and J. L. Gauvain, “Connectionist language modeling for large vocabulary continuous speech recognition,” in *Proc. ICASSP*, 2002.
- [6] I. Bulyko et al., “Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures,” in *Proc. HLT-NAACL*, 2003, vol. Comp., pp. 7–9.
- [7] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” in *Proc. ACL*, 1998.
- [8] D. Biber, *Variation Across Speech and Writing*, Cambridge University Press, 1988.
- [9] W. J. M. Levelt, “Monitoring and self-repair in speech,” *Cognition*, vol. 14, pp. 41–104, 1983.
- [10] J.M. Van Thong et al., “Real-world audio indexing systems,” in *Proc. ICASSP*, 2005.
- [11] W. Byrne et al., “Automatic recognition of spontaneous speech for access to multilingual oral history archives,” *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 4, pp. 420–435, 2004.
- [12] D. Reynolds and P. Torres-Carrasquillo, “Approaches and applications of speech diarization,” in *Proc. ICASSP*, 2005.
- [13] M. Snover, R. Schwartz, B. Dorr, and J. Makhoul, “RT-S: Surface rich transcription scoring, methodology, and initial results,” in *Proc. DARPA Rich Transcription Workshop*, 2004.
- [14] Y. Liu and et al., “Structural metadata research in the EARS program,” in *Proc. ICASSP*, 2005.
- [15] M. Lease et al., “Parsing and its applications for conversational speech,” in *Proc. ICASSP*, 2005.
- [16] P. Heeman and J. Allen, “Speech repairs, intonational phrases and discourse markers: Modeling speakers’ utterances in spoken dialogue,” *Computational Linguistics*, vol. 25, pp. 527–571, 1999.
- [17] R. Kompe, *Prosody in Speech Understanding Systems*, Springer-Verlag, 1996.
- [18] J. Ang, Y. Liu, and E. Shriberg, “Automatic dialog act segmentation and classification in multiparty meetings,” in *Proc. ICASSP*, 2005.
- [19] G. Ji and J. Bilmes, “Dialog act tagging using graphical models,” in *Proc. ICASSP*, 2005.
- [20] M. Galley et al., “Discourse segmentation of multi-party conversation,” in *Proc. ACL*, 2003, pp. 562–569.
- [21] A. Clark and A. Popescu-Belis, “Multilevel dialogue act tags,” in *Proc. SIGDIAL*, 2004, pp. 163–170.
- [22] R. Cowie et al., “Emotion recognition in human-computer interaction,” *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.
- [23] S. Kajarekar et al., “SRI’s 2004 NIST speaker recognition evaluation system,” in *Proc. ICASSP*, 2005.
- [24] D. Hindle, “Deterministic parsing of syntactic nonfluencies,” in *Proc. ACL*, 1983, pp. 123–128.
- [25] E. Charniak and M. Johnson, “Edit detection and parsing for transcribed speech,” in *Proc. NAACL*, 2001, pp. 118–126.
- [26] M. Gregory et al., “Sentence-internal prosody does not help parsing the way punctuation does,” in *Proc. NAACL*, 2004, pp. 81–88.
- [27] J. Kahn et al., “Parsing conversational speech using enhanced segmentation,” in *Proc. HLT-NAACL*, 2004, pp. 125–128.
- [28] E. Nöth et al., “Verbmobil: The use of prosody in the linguistic components of a speech understanding system,” *IEEE Trans. SAP*, vol. 8, no. 5, pp. 519–532, 2000.
- [29] L. Mangu et al., “Finding consensus in speech recognition: word error minimization and other applications of confusion networks,” *Computer Speech and Language*, pp. 373–400, 2000.
- [30] D. Palmer and M. Ostendorf, “Improving information extraction by modeling errors in ASR output,” in *Proc. HLT Workshop*, 2001, pp. 156–160.
- [31] G. Chung, “Automatically incorporating unknown words in Jupiter,” in *Proc. ICSLP*, 2000, vol. IV, pp. 520–523.
- [32] B. Logan and JM. Van Thong, “Confusion-based query expansion for OOV in spoken document retrieval,” in *Proc. ICSLP*, 2002.
- [33] D. Palmer and M. Ostendorf, “Improving out-of-vocabulary name resolution,” *Computer Speech and Language*, vol. 19, no. 1, pp. 107–128, 2005.
- [34] S.-F. Chang et al., “Combining text and audio-visual features in video indexing,” in *Proc. ICASSP*, 2005.
- [35] D. Munteanu et al., “Improved machine translation performance via parallel extraction from comparable corpora,” in *Proc. HLT-NAACL*, 2004, pp. 265–272.
- [36] M. Lapata and F. Keller, “The Web as a baseline: evaluating the performance of unsupervised Web-based models for a range of NLP tasks,” in *Proc. HLT-NAACL*, 2004, pp. 121–128.
- [37] K. Papineni et al., “Bleu: A method for automatic evaluation of machine translation,” in *Proc. ACL*, 2004.
- [38] M. Collins, “Discriminative reranking for natural language parsing,” in *Proc. ICML*, 2000, pp. 175–182.
- [39] D. Marcu and K. Knight, “Machine translation in the year 2004,” in *Proc. ICASSP*, 2005.
- [40] L. Ramshaw and R. Weischedel, “Information extraction,” in *Proc. ICASSP*, 2005.
- [41] K. McKeown et al., “From text to speech summarization,” in *Proc. ICASSP*, 2005.
- [42] D. Jones and et al., “Measuring human readability of machine generated text – three case studies in speech recognition and machine translation,” in *Proc. ICASSP*, 2005.
- [43] D. Palmer, “User-centered evaluation for machine translation of spoken language,” in *Proc. ICASSP*, 2005.
- [44] Y. Gao et al., “Portability challenges in developing interactive dialog systems,” in *Proc. ICASSP*, 2005.