Junlan Feng, Bhuvana Ramabhadran,
John H.L. Hansen, and
Jason D. Williams

Trends in Speech and Language Processing

Speech and language processing covers a broad range of research topics at the intersection of signal processing and language, including speech recognition, text-to-speech synthesis (TTS), spoken language understanding, translation, spoken dialog management, speech indexing, information extraction, speaker and language recognition, and others. At ICASSP, almost 25% of all papers submitted and presented are in these areas.

Speech and language processing is gaining wider acceptance within businesses, with hundreds of deployments from some of the world's largest corporations, numerous small specialized companies, and government and military institutions. In addition, the study of speech and language processing expands our knowledge of human spoken language.

This article summarizes current trends in speech and language processing, adapted from the ICASSP 2011 session "Trends in Speech and Spoken Language Processing (SLP)." Given the space restrictions of the column, it is not exhaustive, but rather includes selected trends that, in our opinion, will shape the future of the field.

SPEECH RECOGNITION

Automated speech recognition (ASR) is the foundation of many speech and language processing applications. ASR technology includes signal processing, optimization, machine learning, and statistical techniques to model human speech and understanding. Benefitting from advances made from numerous research programs funded by the United States, the European Union, and Asian agencies, a broad spectrum of real-world applications have emerged, including simple transcription of news bulletins, multichannel conversations over noisy telephone connections, information retrieval and translation. Over one third of the papers in speech and language processing presented at ICASSP 2011 addressed key challenges in ASR. Broadly speaking, ASR can be separated into acoustic modeling, language modeling, and search.

ACOUSTIC MODELING

An interesting trend is the increasing use of machine learning techniques for acoustic modeling. New modeling paradigms that capture the higher-order hidden structure inherent in the data, such as deep belief networks that employ restricted Boltzmann machines (RBMs), and continuous random fields (CRFs) are beginning to complement and, in some cases, outperform hidden Markov model (HMM)-based models that have dominated the field for many decades. Efforts are also being made to scale these techniques to large volumes of training data, which is increasingly available from real-world applications such as "voice search" (searching the Internet by voice). For example, progress has been made for better acoustic unit selection, training criteria that directly minimize the errors an ASR system makes, noise-robust features, rapid adaptation of models to achieve good performance with few seconds of data and efficient parameter estimation via kernels to handle nonconvex optimization. Key challenges remain, whether it is in the transfer of promising approaches such as compressive sensing, sparse representations, constrained subspace modeling, or in developing new methods for unsupervised training with unlabeled data.

LANGUAGE MODELING

Recent developments include complex models from the exponential family, such as Model M (a class-based exponential language model), long-span models such as neural network-based models, and discriminatively trained language models. In part due to advances in distributed computing, these methods are yielding higher performance over conventional N-gram models on various tasks and languages. ICASSP 2011 saw several contributions to core algorithms that scale well to billions of words in the training data, distributed training paradigms, and selection of features including morphological, syntactic, and semantic features.

SEARCH

Efficient decoding (search) strategies have been developed in conjunction with new acoustic and language models so that they may be used for recognition. Enhancements to finite-state transducers and hill-climbing methods on speech lattices show promise.

TEXT-TO-SPEECH SYNTHESIS

Speech synthesis is the process of enabling machines to speak naturally like humans. Two broad classes of TTS systems exist today: concatenative TTS, which is the process of stitching together context-dependent units of speech subject to duration and pitch constraints; and parameterized TTS, where a HMM is used to generate speech given a set of acoustic and linguistic features. Speech parameters such as spectrum, excitation, and duration depend on a variety of contextual factors such as phoneme identities, accent, and parts of speech. In the HMM-based speech synthesis system, context-dependent models are generally used to capture these contextual factors. Recent trends have focused on improving the underlying parameterization and reconstruction algorithms in HMM-based TTS systems with lesser emphasis on concatenative approaches. However, hybrid systems that combine these two approaches are very promising. One possible reason for this trend is that embedded HMM-based TTS systems require a small footprint and are an attractive solution for low-resource and fixed-point only arithmetic devices (cell phones). TTS is also an area where machine learning techniques

Digital Object Identifier 10.1109/MSP.2011.943131
Date of publication: 8 December 2011

are being employed for pitch accent/prosody (the perceived *pitch* of human speech) prediction with limited success. Techniques to generate emotional speech, voice conversion methods, automated part of speech tagging that is optimized for speech synthesis and adaptation and modeling prosody were some of the challenges addressed by researchers. In speech-to-speech (S2S) translation, the naturalness and intelligibility of the synthesized speech is strongly affected by the fluency of the translated sentences. This is an emerging area with emphasis on better integration of recognition, translation, and synthesis technologies.

SPEAKER RECOGNITION, VERIFICATION, AND DIARIZATION

There were more than 40 papers presented at ICASSP 2011 on speaker recognition and verification. A few recently proposed features and algorithms are becoming mainstream. "I-Vector," which was first proposed in 2008 by Dehak et al., has offered the state-of-the-art performance on many speaker recognition and verification tasks. I-vector is a compact representation of a speaker utterance extracted from a low-dimensional total variability subspace. The basic idea is that the I-vector extractor converts a sequence of frame-based features to a single low-dimensional vector representing the whole utterance. Linear discriminant analysis is the commonly used technique to score the I-vector so that it can minimize within-speaker variability and maximize the between-speaker variability. Another commonly accepted approach in speaker recognition is system fusion. Many systems participating in the National Institute of Standards and Technology (NIST) 2010 Speaker Recognition Evaluation (SRE) have shown that the combined system, which combines speaker recognition scores from multiple systems, achieved significant performance gain compared to each individual system.

Speaker diarization infers who spoke when in an audio stream, meeting, or video. The proposed approaches include top-down and bottom-up clustering, speaker role N-gram models, binary keys, and the information bottleneck-based approach.

SPEECH ENHANCEMENT AND ANALYSIS

Speech enhancement continued to be a very active topic at ICASSP 2011. The research problems include echo cancellation, residual echo reduction, noise reduction, dereverberation, artificial bandwidth extension, and speaker/source separation. In terms of applications, speech enhancement has recently been applied for digital hearing aids and music processing. Another hot topic is speech enhancement using distributed microphones, where the subareas of speech enhancement, particularly, source localization, source separation, and compressed sensing meet. Major progress has been made in speaker/source separation using binary spectral masks and sophisticated estimation techniques.

Speech analysis investigates a wide range of problems such as utterance level emotion detection, duration modeling for ASR, pitch frequency estimation, measuring impact of varying types of noises on ASR, phonetic segmentation, quantifying perturbations, hearing loss simulation, and time-frequency analysis. Noise-robust processing algorithms have become more important. Two of the examples are F0 estimation in noise and voice activity detection in noise. Researchers in speech analysis are also looking beyond the spectrograms such as using auditory representations. The trending approaches on speech analysis include singularity exponents (SEs), conditional random fields, phase-locked loops (PLLs) as well as subglottal resonances detection to help in speaker normalization efforts.

SPOKEN LANGUAGE PROCESSING

In the past few years, the number of papers on spoken language processing at ICASSP has consistently grown. The problems researchers focus on range from well-established topics such as S2S translation, lexical modeling, and speech document classification/summarization to fairly new topics such as speakers' cognition load detection, language identification, multilingual/multichannel processing, and spoken term detection. Many of the advances on these topics have been motivated by recent industrial needs as well as a number of evaluation series. Speech applications on mobile devices, speech recognition for new languages, and translation are some of the top industrial demands. Evaluation examples include DARPA Global Autonomous Language Exploitation (GALE) Program, NIST spoken term detection evaluation, and NIST language recognition evaluation.

In terms of techniques, conditional random fields, logistic regression, and latent Dirichlet allocation (LDA) are increasingly applied for classification and summarization tasks. Another trend is to tightly couple speech processing and language understanding for better overall application performance. A combination of acoustic features and linguistic features has shown improvements on tasks such as language identification and voice search.

STATISTICAL SPOKEN DIALOG SYSTEMS

Spoken dialog systems—computer agents that interact with people using spoken language—are an important application of spoken language technology. In this area, an important trend has been toward statistical spoken dialog systems, which encompass two key ideas. The first idea is to track a distribution over all possible dialog states, rather than a single hypothesis. The result is dialog systems that are more robust to errors. The second idea is to choose actions using reinforcement learning, rather than following a hand-crafted design. The result is a dialog system with a more fine-grained dialog plan that more often yields successful dialogs. The marriage of these two ideas leads to casting a dialog system as a partially observable Markov decision process (POMDP), a formalism studied widely in the operations research, planning, and machine learning communities. Recent studies have shown that casting dialog systems as POMDPs increase performance in the lab; this year, statistical dialog systems have been deployed "in the wild" to real users for the first time and are showing positive results.

LOOKING AHEAD: GRAND CHALLENGES IN SPEECH AND LANGUAGE PROCESSING

Thanks to a host of advancements over the past 30 years, many speech and language technologies are now a part of daily life. However, a number of impediments remain to making these technologies ubiquitous. In light of the increasingly mobile and socially connected population, core challenges include robustness to noise, communication channels, handsets, and microphone mismatch. Other challenges include the ability to support the world's range of languages and dialects in speech technology.

Looking beyond these obstacles for general use, it is interesting to contemplate grand challenges for speech and language processing. Here we borrow the term *grand challenges* from the U.S. National Academy of Engineering, who has suggested challenges such as producing energy from fusion, securing cyberspace, reverse-engineering the brain, and making solar energy economical. Here we suggest three grand challenges for speech and language processing.

MULTILINGUAL SPEECH-TO-SPEECH TRANSLATION

Recently, a number of efforts have emerged and have demonstrated effective S2S translation. With more than 6,000 languages spoken in the world, the ability to reduce communication barriers between humans could 1) help reduce differences between peoples where military or social conflicts might arise; 2) provide more effective rapid response by emergency and care givers in times of natural disasters; 3) help encourage closer cooperation in science and engineering advancements; and 4) simply help those traveling to new countries better interact with others and appreciate cultures. Mobile technology in the form of cell phones (Android, iPhone, etc.) has enabled improved computing support for mobile communication devices. Technology breakthroughs have taken place in this domain and are being used in real-world scenarios. It should be clear that seamless S2S translation is something that would benefit all.

Slides

SPEECH RECOGNITION FOR ALL LANGUAGES

Many of the languages and dialects in the world have no written form; a growing number are considered "dying languages" because the numbers of speakers are dwindling. Wikipedia lists ten languages spoken by more than 100 million speakers and another 12 with between 50–100 million speakers. It is estimated that 330 languages are spoken by more than 1 million people. However, if one considers how many languages enjoy the most effective working speech recognition platforms, the number might be less than 30. As such, there is a cultural, economic, and societal need to see speech recognition as well as various forms of language technology move to new and underresearched languages. In the United States, Intelligence Advanced Research Projects Activity (IARPA) recently announced their goals to focus on this topic in the BABEL program, which will develop agile and robust speech recognition technology that can be rapidly applied to any human language. Also, organizations such as The Computer Assisted Language Instruction Consortium (CALICO) have focused on improving language instruction via interactive computer systems that address text, speech pronunciation (i.e., speech recognition), and perception (i.e., proper intonation and pronunciation). Advancements here clearly would represent one of the core grand challenges in speech and language technology.

THOUGHT-TO-SPEECH SIGNAL PRODUCTION/SPEAKER DEPENDENT SYNTHESIS

One area that has challenged speech scientists is the ability to tap directly into the thought process of the brain to translate concepts/ideas into a speech signal. Mapping the speech communication thought process from Wernicke's area to Broca's area and then to the motor cortex is not an easy task, along with the speech articulators in the motor cortex is not an easy task. For those individuals who suffer from permanent paralysis/inability to vocalize any speech, the prospect of implanting a microelectrode array into the language portion of the brain and tapping into the neural activity directly [i.e., a brain machine interface (BMI)] would allow for someone to "think" of something they would want to say, and that information would be sensed and transmitted wirelessly to an external speech synthesis engine where artificial speech is produced. For subjects who know they will lose their ability to speak (e.g., pending surgery on the larynx, etc.), collecting sufficient speech content prior to surgery would allow an individual to maintain their voice (Robert Ebert, the American movie critic, is an example of how saving prior speech can help restore one's voice after severe health recovery). Personalized synthesis methods therefore represent another domain.

In summary, for speech and language processing, years of research progress now underpins numerous commercial applications, and these advancements are clearly migrating into everyday use. Even so, a number of important and interesting challenges not only remain but have been created as these technologies migrate into real-world applications, and we look forward to seeing further advances at conferences such as IEEE ICASSP and workshops such as the IEEE Automatic Speech Recognition and Understanding Workshop.

AUTHORS

Junlan Feng (junlan@research.att.com) is a principal member of technical staff at AT&T Labs Research.

Bhuvana Ramabhadran (bhuvana@us.ibm.com) is a research staff member and the manager of the Speech Transcription and Synthesis Research Group, IBM T.J. Watson Center, Yorktown Heights, New York.

John H.L. Hansen (john.hansen@utdallas.edu) is the department head and professor at The University of Texas at Dallas, Erik Jonsson School of Engineering and Computer Science, where he holds the Distinguished University Chair in Telecommunications Engineering and a joint appointment as professor in the School of Behavioral and Brain Sciences (Speech and Hearing).

Video

Jason D. Williams (jdw@research.att.com) is a principal member of technical staff at AT&T Labs Research.