Towards Unsupervised Speech Processing

James Glass

MIT Computer Science and Artificial Intelligence Laboratory,
32 Vassar St., Cambridge, MA 02139, USA

ABSTRACT

The development of an automatic speech recognizer is typically a highly supervised process involving the specification of phonetic inventories, lexicons, acoustic and language models, and requiring annotated training corpora consisting of parallel speech and text data. Although some model parameters may be modified via adaptation, the overall structure of the speech recognizer usually remains relatively static. While this approach has been effective for problems where there is adequate human expertise, and labelled corpora are available, it is challenged by less-supervised or unsupervised scenarios. It also contrasts sharply with human speech processing where learning is an inherent ability.

In this paper, three alternative scenarios for speech recognition "training" are described, each requiring decreasing amounts of human expertise and annotated resources, and increasing amounts of unsupervised learning. A speech deciphering challenge is then suggested whereby speech recognizers must learn sub-word inventories and word pronunciations from unannotated speech, supplemented with only non-parallel text resources. It is argued that such a capability will help alleviate the language barrier that currently limits the scope of speech recognition capabilities around the world, and empower speech recognizers to continually learn and evolve through use.

1. INTRODUCTION

The field of automatic speech recognition (ASR) has made tremendous advances over the last thirty years. During this time, ASR technology has consolidated around a set of components that are illustrated in Figure 1. These include 1) a signal processor to generate a representation of the speech signal in terms of a sequence of acoustic observations, 2) acoustic models that provide evidence as to the likely sound sequence in the waveform, 3) a pronunciation lexicon to provide a mapping between vocabulary words and their associated sub-word unit realization, and 4) a language model to provide guidance as to likely word order. The latter three components are typically incorporated into a search to find the most likely sequence of words given the acoustic observations.

Popular representations and modeling methods for modern ASR systems include Mel-Frequency Cepstral Coefficients [1], Gaussian mixture models, hidden Markov models [2], n-gram language models, weighted finite-state transducers [3], Viterbi and N-best search [4], etc. Many of these concepts, including the notion of representing sub-word units and language model constraints as a search graph [5], and the probabilistic formulation for hypothesizing words [6], were first applied to ASR in the 1970s [7], became adopted on a wide-scale in the 1980s [8], and have become standard ASR practice since that time [9].

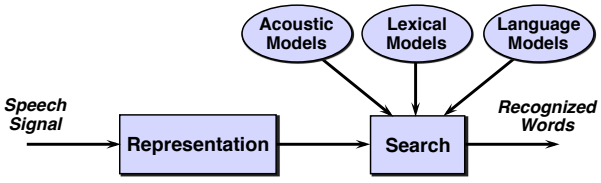

Fig. 1. Automatic speech recognition components.

One property of modern speech recognition technology is that it typically incorporates expert knowledge of a language, and undergoes a heavily supervised training process. Linguistic expertise is often provided in the form of the pronunciation lexicon and associated set of sub-word units, which typically correspond to phoneme-like inventories. Supervised training occurs through the use of large speech corpora, where each speech recording is usually associated with a parallel word-level transcription. When these linguistic resources are available, the training process, either using maximum likelihood or discriminative techniques, is well established [9].

The supervised training paradigm has served the ASR community well. As computation, memory, and the size of annotated corpora have increased, the stochastic models and training techniques have become more sophisticated, resulting in a relatively continuous reduction of word error rates on increasingly challenging tasks [10]. The techniques and methodologies that have been developed in the ASR community over the years have arguably influenced other research areas as well.

Although there is considerable active research in developing better signal representations, acoustic, lexical, and language models, as well as faster search algorithms, it is also worthwhile considering that there are alternative scenarios for processing speech than the current well-established framework. While the supervised training approach has been effective for scenarios where there is adequate human expertise and labelled corpora, it is challenged by less-supervised or even unsupervised scenarios. These dual requirements, and their associated cost, are a big part of the reason that a relatively small fraction of the world's languages have a functional ASR capabil-

978-1-4673-0382-8/12/$31.00 ©2012 IEEE

ity [11]. These constraints are in sharp contrast with human speech processing requirements where learning is an inherent ability [12, 13]. All humans process vast quantities of unannotated speech and manage to learn phonetic inventories, word boundaries, etc., and can use these abilities to acquire new words [14, 15, 16]. Why can't ASR technology have similar capabilities?

In the following section, three alternative scenarios for speech recognition learning are described that require fewer annotated resources and human expertise. An unsupervised speech recognition challenge is then presented that requires learning phonetic inventories, and word pronunciations from non-parallel speech and text sources.

2. UNSUPERVISED SPEECH PROCESSING

As illustrated in Figure 2, there is a range of scenarios for processing speech that can be characterized by decreasing amounts of human expertise and supervised intervention, and a corresponding increase in unsupervised learning and technical difficulty.

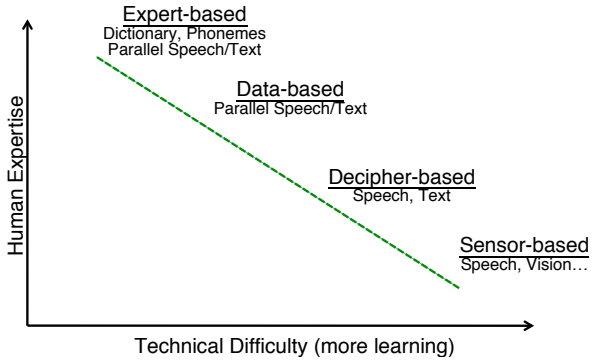

Fig. 2. Potential ASR learning scenarios.

2.1. Expert-based Speech Processing

At one end of the spectrum is the conventional "expert-based" approach that uses a pronunciation dictionary as the cornerstone to anchor the mapping between sound sequences and word sequences, and exploits annotated training data with parallel speech and text to learn model parameters. This point represents most ASR research that has been conducted since the late 1980s [17], when sub-word hidden Markov models (HMMs) became the dominant ASR paradigm [18].

2.2. Data-based Speech Processing

If the pronunciation dictionary and linguistic units are not provided by an expert, then the issue of learning the dictionary and associated inventory of units automatically must be addressed. For languages that have straightforward letter-to-sound mappings, a grapheme-based approach has been shown to be effective [19]. For languages where this is not the case, however, the challenge posed by this "data-based" scenario is whether pronunciations can be learned automatically from annotated data, and whether a data-driven set of units can outperform a set of more conventional linguistically specified units. Note that it is somewhat ironic that, although all other parameters of most ASR systems are learned from data, the conventional pronunciation dictionary is still specified by experts. It is not unreasonable to expect that, ultimately, automatically learned units could out-perform manually specified units.

As an example of research headed in this direction, we have been exploring the use of a pronunciation mixture model, that can learn pronunciations for an entire lexicon from an initial letter-to-sound (L2S) model and annotated speech data [20]. If such a capability can be developed without an initial L2S model - perhaps by iteratively learning sub-word units and their associated n-gram statistics, then reasonable ASR capability should be achievable for languages where annotated data is available.

The data-based scenario has several variations that include combinations of annotated and unannotated data, as well as approaches that explore a combination of annotated data from different languages in order to more easily learn the ASR parameters of a new language [21]. There are many real-world scenarios where a limited amount of annotated data may be available along with much larger inventories of unannotated data and there has been active research in this area [22], as well as in active learning [23]. Another area of research related to this scenario is one whereby low-cost human knowledge can be incorporated into the learning process. For example, there has been much recent activity on using crowdsourcing techniques for collecting and annotating speech data [24, 25].

2.3. Decipher-based Speech Processing

A major break from conventional ASR training would be achieved by techniques that are able to learn from unannotated speech combined with non-parallel text data. Although the text data may be available to learn vocabularies and language models, the determination of what words occur where will need to be inferred from the data.

In the ASR community, there has been recent activity on learning from unannotated speech - a scenario sometimes referred to as the zero resource scenario. Researchers have shown, for example, that it is possible to identify word-like patterns in the speech signal by looking for re-occurring sound sequences [26, 27, 28, 29, 30, 31]. This work is related in spirit to work on 'motif' discovery in the data mining community [32]. Automatically discovered speech patterns have been used for a variety of applications, including spoken query retrieval [33], topic segmentation [34], topic classification [35, 36], and unit learning [37]. There has also been research directed towards automatically finding an appropriate set of sub-word units from speech data alone, sometimes referred to as self-organizing units (SOUs) [38, 39].

A logical next step in these latter endeavors is to make the connection with non-parallel text data, in order to learn a pronunciation dictionary automatically. Theoretically at least, knowledge of the lexicon should provide constraint

as to the inventory of linguistic units for a language. While this may seem like a daunting problem, it may be viewed as a kind of deciphering task which has received some recent attention in the machine translation community for non-parallel text corpora [40, 41], and it is possible that such approaches may be useful for the speech "decipher-based" scenario as well.

2.4. Sensor-based Speech Processing

At the extreme end of unsupervised speech processing lies an area which most closely matches that of human spoken language acquisition. In this "sensor-based" speech processing scenario, incoming speech signals might be paired with other sensory inputs such as vision. A natural application for this capability is one that involves human-machine interaction. For example, this scenario might be appropriate for a robot in a new environment learning language capability through spoken interactions. While this may appear to be a futuristic, unrealistic, scenario, there has been research in this area that indicate some ability to jointly learn linguistic and perceptual models of semantic concepts [42]. Certainly, for future interaction with robots, it will be desirable to be able to teach the robot new concepts through spoken interaction.

3. A SPEECH DECIPHERING CHALLENGE

Of the four speech scenarios described in the previous section, the decipher-based scenario is the most obvious candidate to serve as a challenge to the speech community. The expert-based scenario is the current one that we have been using for the past several decades, and which continues to dominate current ASR research. The data-based scenario has been at least partially addressed by grapheme-based approaches, and is, arguably, on the verge of being solved by virtue of automated dictionary learning methods. The sensor-based approach is more in the realm of human-robot interaction, and I believe research in speech, language, vision, and robotics will result in advances in this area in the coming years.

The decipher-based scenario can be considered an unsupervised challenge for the speech community. Given a corpus of unannotated speech, and an independent corpus of text data, can we develop techniques that will automatically 1) learn an appropriate set of sub-word units, 2) make the connections between "words" in the text data and the spoken realizations, and 3) learn a pronunciation dictionary of these words in terms of the learned sub-word units? One way to evaluate such a capability would be to measure the transcription accuracy of a test data from the same language, although other measures, such as spoken term detection, could also be used. There have also been recent efforts on unsupervised testing as well [43].

If such a capability could be achieved then arguably we will have broken the language barrier, for many more languages than have current ASR capability. We will also make headway towards sensor-based scenarios where text data are incorporated. Note that there is no reason this task could not incorporate human-level feedback about whether certain hypothesized words are appropriate (imagine an Amazon Mechanical Turk task, for example). Note also that there is no reason that linguistic information about the nature of the inventories of speech sounds in the languages of the world, nor of other generic phonological structures at the syllable, word level, etc., could not be incorporated as *a priori* information. The challenge, as it has always been, is to find a good stochastic model within which to incorporate these constraints.

4. CONCLUSIONS

While the four speech processing scenarios described here might seem increasingly outlandish and impractical, I believe that exploring methods that incorporate additional unsupervised learning and require less supervised training will allow us to break through the language barrier that we face due to the sheer cost and effort of creating large annotated corpora and associated dictionaries for new languages. Moreover, any unsupervised learning methods that are effective are likely to benefit more traditional ASR scenarios as well. These techniques will enable ASR technology to become more adaptable as learning becomes a more intrinsic capability.

In summary, a move towards less supervised or unsupervised speech processing poses a research challenge. Due to recent advances in machine learning methods, arguably, the time is ripe to make progress in this area. The question is, how long will it take the speech community to make headway? Let the adventure begin!

5. ACKNOWLEDGEMENTS

The author would like to thank Hung-An Chang, Ekapol Chuangsuwanich, Ian McGraw, Stephanie Seneff, T.J. Hazen, and Yaodong Zhang for their helpful suggestions.

6. REFERENCES

[1] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Rec. in Cont. Spoken Sents.," *IEEE Trans. ASSP,* 28, 1980.

[2] J. Baker, "The DRAGON System – An Overview," *IEEE Trans. ASSP,* 23, 1975.

[3] F. Pereira, M. Riley, R. Sproat, "Weighted Rational Transductions and their Application to Human Language Processing," *Proc. HLT*, San Francisco, 1994.

[4] F. Soong and E. Huang, "A Tree-Trellis Based Fast Search for Finding the N-Best Sentence Hypothesis in Cont. Speech Rec.," *Proc. ICASSP,* Toronto, 1991.

[5] B. Lowerre, "Dynamic Speaker Adaptation in the Harpy Speech Recognition System," *Proc. ICASSP,* Hartford, 1977.

[6] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE,* 64, 1976.

[7] W. Lea, "Trends in Speech Recognition," *Prentice-Hall*, 1980.

[8] A. Waibel and K.F. Lee, "Readings in Speech Recognition," *Morgan Kaufman*, 1990.

[9] X. Huang, A. Acero, and H.W. Hon, "Spoken Language Processing," *Prentice-Hall*, 2001.

[10] A. Martin and J. Garofolo, "NIST Speech Processing Evaluations: LVCSR, Speaker Recognition, Language Recognition," *Proc. SAFE*, Washington, 2007.

[11] K. Predoca, "Non-mainstream Languages and Speech Recognition: Some Challenges," *CALICO Journal,* 21(2), 2004.

[12] L. Bloomfield, "Language," *Holt*, New York, 1933.

[13] N. Chomsky, "Knowledge of Language: Its Nature, Origin, and Use," *Praeger*, New York, 1986.

[14] P. Jusczyk, "The discovery of spoken language," *MIT Press,*, Cambridge, 1997.

[15] S. Pinker, "The Language Instinct," *William Morrow and Company*, New York, 1994.

[16] J. Saffran, "Constraints on Statistical Language Learning," *J. Memory and Language*, 47, 2002.

[17] K.F. Lee, "Context Dependent Phonetic HMMs for Cont. Speech Rec.," *IEEE Trans. ASSP,* 38, 1990.

[18] J. Baker et al., "Research Developments and Directions in Speech Recognition and Understanding," *IEEE Signal Processing Magazine,* 2009.

[19] M. Killer, S. Stuker, and T. Schultz, "Grapheme Based Speech Recognition," *Proc. Eurospeech,* Geneva, 2003.

[20] I. Badr, I. McGraw, and J. Glass, "Pronunciation Learning from Continuous Speech," *Proc. Interspeech,* Florence, 2011.

[21] T. Schultz and K. Kirchhoff, "Multilingual Speech Processing," *Academic Press,* 2006.

[22] S. Novotney, R. Schwartz, and J. Ma, "Unsupervised Acoustic and Language Model Training with Small Amounts of Labelled Data," *Proc. ICASSP,* 2009.

[23] G. Riccardi and D. Hakkani-Tür, "Active and Unsupervised Learning for ASR," *Proc. Eurospeech,* 2003.

[24] I. McGraw, C. Lee, L. Hetherington, S. Seneff and J. Glass, "Collecting Voices from the Cloud," *Proc. LREC*, Malta, 2010.

[25] S. Novotney and C. Callison-Burch, "Cheap, Fast, and Good Enough: Automatic Speech Rec. with Non-Expert Transcription," *Proc. ICASSP,* 2010.

[26] A. Park and J. Glass, "Towards Unsupervised Pattern Discovery in Speech," *Proc. ASRU,*, San Juan, 2005.

[27] A. Park and J. Glass, "Unsupervised Pattern Discovery in Speech," *IEEE Trans. ASLP,*, 16(1), 2008.

[28] A. Muscariello, G. Gravier, F. Bimbot, "Audio keyword extraction by unsupervised word discovery," *Proc. Interspeech,* Brighton, 2009.

[29] Y. Zhang and J. Glass, "Towards Multi-Speaker Unsupervised Speech Pattern Discovery," *Proc. ICASSP*, Dallas, 2010.

[30] A. Jansen, K. Church, and H. Hermansky, "Towards Spoken Term Discovery at Scale with Zero Resources," *Proc. Interspeech,* 2010.

[31] M. Siu, H. Gish, S. Lowe, and A. Chan, "Unsupervised Audio Pattern Discovery using HMM-based Self-Organized Units," *Proc. Interspeech,* 2011.

[32] B. Chiu, E. Keogh, and S. Lonardi, "Probabilistic Discovery of Time Series Motifs," *Proc. ICKDDM,* Washington, 2003.

[33] A. Muscariello et al., "Zero-Resource Audio-Only Spoken Term Detect. Based on a Comb. of Template Matching Techniques," *Proc. Interspeech,* 2011.

[34] I. Malioutov, A. Park, R. Barzilay, and J. Glass, "Making Sense of Sound: Unsupervised Topic Segmentation Over Acoustic Input," *Proc. ACL,* 2007.

[35] H. Gish, M. Siu, A. Chan, and W. Belfield, "Unsupervised Training of an HMM-based Speech Rec. System for Topic Class.," *Proc. Interspeech*, 2009.

[36] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, "NLP on Spoken Documents without ASR," *Proc. EMNLP,* Cambridge, 2010.

[37] A. Jansen and K. Church, "Towards Unsupervised Training of Speaker Independent Acoustic Models," *Proc. Interspeech,* Florence, 2011.

[38] M. Siu et al., "Improved Topic Class. and Keyword Discov. Using an HMM-Based Speech Rec. Trained without Supervision," *Proc. Interspeech,* 2010.

[39] C. Lee and J. Glass, "A Nonparametric Bayesian Approach to Acoustic Model Discovery," to appear *ACL*, Jeju, 2012.

[40] B. Synder, R. Barzilay, and K. Knight, "A Statistical Model for Lost Language Decipherment," *Proc. ACL,* 2010.

[41] S. Ravi and K. Knight, "Deciphering Foreign Language," *Proc. ACL,* 2011.

[42] D. Roy and A. Pentland, "Learning words from sights and sounds: A computational model," *Cognitive Science*, 26, 2002.

[43] B. Strope, D. Beeferman, A. Gruenstein, and X. Lei, "Unsupervised Testing Strategies for ASR," *Proc. Interspeech,* Florence, 2011.