An Overview on Automatic Speech Attribute Transcription (ASAT)

Chin-Hui Lee[1], Mark A. Clements[1], Sorin Dusan[2], Eric Fosler-Lussier[3],
Keith Johnson[4], Biing-Hwang Juang[1], Lawrence R. Rabiner[2]
[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[2]Center for Advanced Information Processing, Rutgers University, New Brunswick, NJ, USA
[3]Department of Computer Science and Engineering, Ohio State University, Columbus, OH, USA
[4]Department of Linguistics, University of California, Berkeley, CA, USA
{chl, clements, juang}@ece.gatech.edu, {sdusan, lrr}@caip.rutgers.edu,
keithjohnson@berkeley.edu, fosler@cse.ohio-state.edu

10.21437/Interspeech.2007-509

Abstract

Automatic Speech Attribute Transcription (ASAT), an ITR project sponsored under the NSF grant (IIS-04-27113), is a cross-institute effort involving Georgia Institute of Technology, The Ohio State University, University of California at Berkeley, and Rutgers University. This project approaches speech recognition from a more linguistic perspective: unlike traditional ASR systems, humans detect *acoustic* and *auditory* cues, weigh and combine them to form theories, and then *process* these *cognitive* hypotheses until linguistically and pragmatically consistent speech understanding is achieved. A major goal of the ASAT paradigm is to develop a detection-based approach to automatic speech recognition (ASR) based on attribute detection and knowledge integration. We report on progress of the ASAT project, present a sharable platform for community collaboration, and highlight areas of potential interdisciplinary ASR research.

Index Terms: attributes, events, features, detection, speech recognition, speech attribute transcription, utterance verification

1. Introduction

It has long been postulated that human listeners determine the linguistic identity of sounds based on detected features that exist at various levels of the speech knowledge hierarchy, from acoustics to pragmatics. When comparing state-of-the-art automatic speech recognition (ASR) systems with *human speech recognition* (HSR) [1, 2] it is interesting to note that human beings often perform speech understanding by integrating multiple knowledge sources from the bottom up. Indeed, people do not continuously convert a speech signal into words as an ASR system attempts to do. Instead, they detect *acoustic* and *auditory* cues, weigh them and combine them to form *cognitive* hypotheses, and then *validate* the hypotheses until consistent decisions are reached. The human-based model of speech processing suggests a candidate framework for developing next generation speech processing techniques that have the potential to go beyond the current limitations of existing ASR systems.

The speech signal contains a rich set of information that facilitates human auditory perception and communication, beyond a simple linguistic interpretation of the spoken input. In order to bridge the performance gap between ASR and HSR systems, the narrow notion of speech-to-text in ASR has to be expanded to incorporate all related information "embedded" in speech utterances. This collection of information includes a set of fundamental speech sounds with their linguistic interpretations, a speaker profile encompassing gender, accent, emotional state and other speaker characteristics, the speaking environment etc. Collectively, we call this superset of speech information the *attributes* of speech. They are not only critical for high performance speech recognition but also useful for many other applications, such as speaker recognition, language identification, speech perception, speech synthesis, etc. Based on this set of speech attributes, ASR can be extended to *Automatic Speech Attribute Transcription*, or ASAT, a process that goes beyond the current simple notion of word transcription. ASAT therefore promises to be knowledge-rich and capable of incorporating multiple levels of information in the knowledge hierarchy into attribute detection, evidence verification and integration, i.e. all modules in an ASAT system [3]. Since speech processing in ASAT is highly parallel a collaborative community effort can be built around a common sharable platform to enable a "divide-and-conquer" ASR paradigm that facilitates tight coupling of interdisciplinary studies of speech science and speech processing [4]. A block diagram of the ASAT approach to ASR is shown in Figure 1. The speech signal is first processed by a *bank of speech attribute and event detectors*, followed by *a sequence of event mergers and verifiers*, integrating low-level cues into high-level units and validating evidence to make recognition decisions and identify embedded attributes in speech.

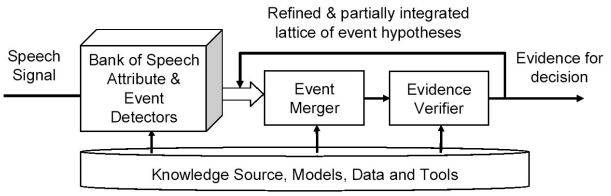

Figure 1 Automatic speech attribute transcription

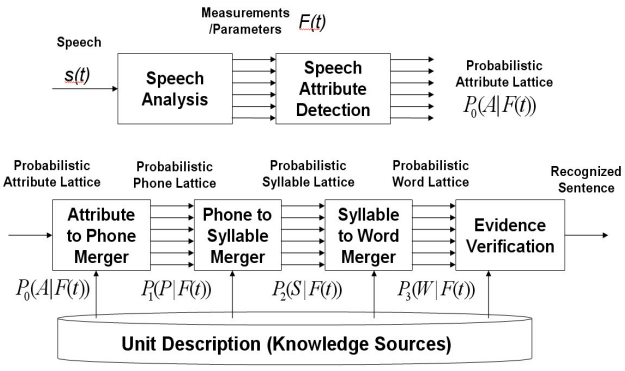

Figure 2 ASAT frontend for analysis and attribute detection, and backend for event merging and evidence verification

Conceptually the ASAT process can be divided into two major components as shown in Figure 2: (1) ASAT frontend processing [5]: performing speech analysis to collect a wide variety of speech parameters, $\{F(t)\}$, at different frame rates, and combining these features to detect speech attributes and form an attribute lattice; and (2) ASAT backend processing: merging and integrating low-level attributes to detect high-level speech units such as syllables, words and sentences, and validating plausible theories based on detected evidence, and pruning unlike hypotheses that may constitute local evidence for making a wrong decision. These two components will be discussed in detail in the following sections.

2. ASAT Frontend Processing

An *event detector* converts an input speech signal *x(t)* into a time series, which describes the level of presence (or *level of activity*) of a particular property (or *attribute*) in the input speech utterance over time. This function can be computed as the *a posteriori* probability of the particular attribute, given the speech signal, within a proper time window, or the *likelihood ratio* (which involves calculation of two likelihoods, one pertaining to the target model and the other the contrast model). The bank of detectors consists of a number of such attribute detectors, each being individually and optimally designed for the detection of a particular property. These properties are often stochastic in nature and are relevant to information needed to perform ASR. One key feature of the detection-based approach is that the outputs of the detectors do not have to be synchronized in time and therefore the system is flexible enough to allow a direct integration of both short-term detectors, e.g., for detection of VOT, and long-term detectors, e.g., for detection of pitch, syllables, and particular word sequences. The conventional frame synchronous constraints of most traditional ASR systems are thus relaxed in the ASAT system.

Speech parameterization techniques have been discussed in many textbooks (e. g., [6]). For ASAT, the parameters can be sample based, such as zero-crossing rate, or frame based, such as mel-frequency cepstral coefficients. Speech analysis can be performed in the temporal domain, providing features such as voice onset time, or in the spectral domain, such as short time spectral energies in different frequency bands. Both long-term and short-term analyses can be compared and contrasted. We have compiled a library of 22 analysis routines in MATLAB code and documented them at the ASAT website [7]. Biologically-inspired and perception-motivated signal analysis are considered as promising parameter extraction directions [8, 9] because the ASAT paradigm supports parameter extraction at different frame rates for designing a range of speech detectors.

Once a collection of speech parameters, $\{F(t)\}$, are obtained, they can be used to perform speech attribute detection which is a critical component in the ASAT paradigm as shown in the upper panel of Figure 2. An example of speech attribute detection is shown in Figure 3, in which three attribute detectors, trained on Mandarin speech for three manner features, namely stop, nasal and vowel, were used to process an English utterance spoken by a non-native Mandarin speaker. In the bottom three panels, detection curves simulating posterior probabilities of stop, nasal and vowel features as a function of time are displayed. It is clear that both stop and nasal manners are correctly detected while all but two speech segments marked "xx" in the bottom panel were correctly detected as vowels due to mispronunciation.

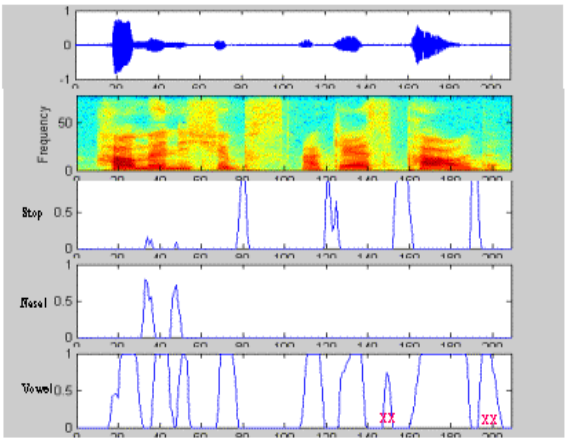

Figure 3 Detectors trained on Mandarin speech and tested on an English utterance, spoken by a non-native speaker

Attributes can be used as cues or landmarks in speech [10] in order to identify "islands of reliability" for making local acoustic and linguistic decisions, such as energy concentration regions and phrase boundaries, without extensive speech modeling. A detection framework was proposed in [11] to discriminate voiced and unvoiced stops using voice onset time for two-pass English letter recognition. Words and key phrases were successfully used as attributes for word recognition [12] to improve understanding of ill-formed utterances.

Fant [13] and Stevens [14] have consistently advocated the approach of detecting and recognizing distinctive features in speech from an acoustic-phonetic perspective. By integrating such acoustic-phonetic knowledge we can design spectrogram reading machines [15]. It is highly instructive to design high-performance speech attribute detectors to realize early acoustic to linguistic mapping directly from the speech signal for ASR purposes. The missing link in utilizing knowledge to recognize speech lies in designing a *bank of "perfect" attribute detectors*. We argue that such deterministic detectors are hard, if not impossible, to realize in practice. These detectors should be stochastic in nature and then the data-driven modeling techniques in state-of-the-art systems could be extended to a bottom-up detection approach to ASR in which *speech event detection* and *linguistic knowledge integration* play key roles.

Stochastic attribute detectors can be frame-based, such as computing scores on a frame-by-frame basis using an ensemble of artificial neural networks (ANNs) for detecting manner and place of articulation [16], or segment-based using hidden Markov models (HMMs) [17]. Such detectors can also be signal-based [18] and/or hierarchical [19]. They can be added to existing state-of-the-art speech recognition [20, 21] and speaker verification [22] algorithms to improve the capabilities of current systems. Discriminative training can also be directly adopted to improve detector model separation [23].

3. ASAT Backend Processing

Another critical component in the ASAT paradigm is the backend processing shown in the bottom panel in Figure 2. An event merger takes the set of detected lower-level events as input

and attempts to infer the presence of higher-level units (e.g., a phone or a word) which are then validated by the evidence verifier to produce a refined and partially integrated lattice of event hypotheses. Such a lattice of hypotheses is then fed back for further event merger and knowledge integration. This iterative information fusion process always uses the original event activity functions as the raw evidence. A terminating strategy can be instituted by utilizing all the supported attributes. The procedure produces the evidences needed for a final decision, including a recognized sentence.

The proposed merging approach, when applied to auditory processing, attempts to simulate the human auditory process by assuming that speech is first converted to a collection of auditory response patterns (via attribute detection), each representing a probabilistic activity level of a particular acoustic-phonetic event of interest (shown in the A1 module in Figure 4). Detection of the next level of events or evidence, such as phones, is accomplished by combining relevant features from A1 (as shown in the A2 module in Figure 4). Each activity function can be modeled by a corresponding neural system. The ANN framework provides a convenient tool to model neuron combinations. Feed-forward neural networks have been used to encode and decode temporal information. Recurrent neural networks have also been used to provide feedback loops to simulate neural processing. Simulating perception of temporal events is of particular interest in auditory perception of speech.

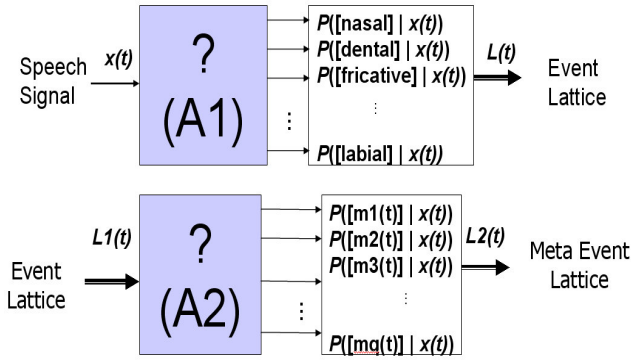

Figure 4 Merging attributes into phones and meta-events

The processing of attribute merging shown in the bottom panel in Figure 4 can be interpreted as a form of probabilistic parsing of event lattices. New techniques are needed to accomplish this form of lattice parsing. An example of combining phonetic attributes for phone recognition is through conditional random fields [24, 25], which allow for the integration of overlapping, redundant cues.

Event verification, a critical ASAT component, is often formulated as a statistical hypothesis testing problem. There are several techniques discussed in the literature for designing optimal tests, if the distributions of the null and alternative hypotheses are known exactly. However, for most practical verification problems, we use a set of training data to estimate the distributions of these competing hypotheses. Utterance verification has been studied extensively [26] for rejecting unlikely speech theories. Use of generalized log likelihood ratio (GLLR) was recently proposed as a way to measure separation between competing hypotheses [27].

4. Unit Description and Knowledge Integration

As shown in Figure 1 and the A2 processing module in Figure 4, unit description is fundamental to knowledge-based processing in ASAT. Such knowledge integration guidelines are widely discussed in classical textbooks [28, 13, 14]. Our goal is to facilitate outside contributions from the ASR and linguistics communities in the ASAT knowledge integration component.

Bottom-up acoustic models can be built hierarchically for attributes, phones, syllables, and words. These outputs can proceed sequentially, as illustrated in Figure 2, or in parallel, as suggested in [29]. This latter approach to knowledge integration is supported by the fact that phonetic information is often spread across adjacent phonemes [30] and that the information about phonetic boundaries can be found in larger segments of speech [31] and it may be useful to increase ASR accuracy [21].

An important consideration in developing descriptions of higher-level units is incorporating the variability of lower-level attributes comprising the higher-level unit. In traditional ASR models, this is incorporated in the variations in the pronunciation and acoustic models; however, in the ASAT paradigm, one can appeal to finer-grained distinctions and develop models of "phonetic ignorance" where particular attributes are discounted (such as voicing for devoiced consonants) [20,30].

5. Collaborative Platform and Evaluation

Collection of language resources and objective evaluation on a large set of real-world utterances are two cornerstones in the advancement of ASR technologies. In order to have rapid progress with the new detection-based ASAT paradigm, a coordinated community effort is needed. In addition to evaluating the word error rate, which is a common practice in ASR benchmarking, we are interested in the performance of the detectors for both low-level attributes and high-level evidence. To facilitate an objective evaluation at the detector level, two scenarios can be considered. The first is a side-by-side, diagnostic comparison between two detectors. The second scenario is similar to ASR system evaluation, that only event-specific detectors are compared. The key to the success of this evaluation methodology is a set of *event-specific evaluation data* that is designed to maximize the coverage of the target attributes with testing on different contexts. A *library of attribute-specific detectors* is also essential. Since the tests are done at the attribute and evidence levels, they should offer lots of diagnostic information to help with improving detector performances. This library, together with event-specific evaluation data, and evaluation results obtained from competing detectors, will be documented in the ASAT website [7] to facilitate objective evaluation. Some of the subsystem components will also be available from the website to enable ASAT-style cross-site collaboration via this platform.

6. Intermediate Applications and Summary

It is clear that we have a long way to go before we can develop a complete ASAT-based ASR system that is competitive in performance with the state-of-the-art systems. However, we believe that through incorporating knowledge sources into speech modeling and processing, the set of recognized attribute sequences, event lattices, and evidence for decisions provides an instructive collection of diagnostic information, potentially beneficial for improving our understanding of speech, as well as enhancing speech recognition accuracy. As an example, we

found that "knowledge scores" computed with detectors for manner and place of articulation provided a collection of complementary information that can be combined with HMM frame likelihood scores to reduce phone and word errors in rescoring [17, 32]. A knowledge-based pruning strategy was also tested in a detection-based ASR system to reduce false alarms in detection and consequently reduce word errors in ASR [33].

In sum, it is noted that the performance in the ASAT system is "additive", i.e., a better module for a feature will produce better performance for the individual module and other modules related to this attribute, and likely the overall system. To facilitate a community effort to monitor progress we will design a collection of evaluation sets for each attribute. Corresponding performance history will be documented. Everyone is welcome to participate in this effort. We hope to eventually obtain a collection of "best" modules collectively provided by the speech community for a wide range of features, so that they can be collaboratively incorporated into the "best" overall next generation ASR system.

Acknowledgements

This work is supported by the NSF grant, IIS-04-27113. Many students are heavily involved in the ASAT project. Their studies are widely cited in the references to reflect their contributions.

References

[1] J. Allen, "How Do Humans Process and Recognize Speech," *IEEE Trans. Speech and Audio Proc*., Vol. 2, No. 4, pp. 567-577, 1994.

[2] R. Lippmann, "Speech Recognition by Human and Machines," *Speech Communication*, *22*, pp. 1-14, 1997.

[3] C.-H. Lee, "From Knowledge-Ignorant to Knowledge-Rich Modeling: A New Speech Research Paradigm for Next Generation Automatic Speech Recognition," *Proc. ICSLP*, Jeju, South Korea, October 2004.

[4] C.-H. Lee**, "**Back to Speech Science **-** Towards a Collaborative ASR Community of the 21st Century," P. Divenyi, S. Greenberg, and G. Meyer, editors, NATO Science Series on *Dynamics in Speech Production and Perception,* pp. 221-244, IOS Press, 2006.

[5] J. Hou, L. R. Rabiner and S. Dusan, "Automatic Speech Attribute Transcription (ASAT) – the Front End Processor," *Proc. ICASSP*, Toulouse, France, May 2006.

[6] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1976.

[7] http://www.ece.gatech.edu/research/labes/asat.

[8] S. Shamma, "On the role of space and time in auditory processing", *Trends in Cognitive Sciences*, 340-348, 2001.

[9] J. Jeon and B.-H. Juang, "Separation of SNR via dimension expansion in a model of the central auditory system," *Proc. ICASSP*, Toulouse, France, May 2006.

[10] S. A. Liu. "Landmark detection for distinctive feature-based speech recognition." *J. Acoust. Soc. Am*., Vol. 100, No. 5, pp. 3417-3430, Nov. 1996.

[11] P. Niyogi and P. Ramesh, "A Detection Framework for Locating Phonetic Events," *Proc. ICSLP-98*, Sydney, 1998.

[12] T. Kawahara, C.-H. Lee and B.-H. Juang, "Key-Phrase Detection and Verification for Flexible Speech Understanding," *IEEE Trans. on Speech and Audio Proc*., Vol.6, No. 6, pp.558-568, 1998.

[13] G. Fant, *Speech Sounds and Features*, MIT Press, 1973.

[14] K. Stevens, *Acoustic Phonetics*, MIT Press, 1998.

[15] V.W. Zue, "Acoustic-Phonetic Knowledge Representation: Implications from Spectrograms Reading Experiments," *NATO ASI on Speech Recognition*, Bonas, France, 1981.

[16] J. Li, Y. Tsao and C.-H. Lee, "A Study on Knowledge Source Integration for Candidate Rescoring in Automatic Speech Recognition," *Proc. ICASSP*, Philadelphia, 2005.

[17] J. Li and C.-H. Lee, "On Designing and Evaluating Speech Event Detectors," *Proc. InterSpeech*, Lisbon, Sept. 2005.

[18] K. Johnson, "Automatic burst detection in conversational speech," *unpublished manuscript*, 2004.

[19] M. Rajamanohar and E. Fosler-Lussier, "An evaluation of hierarchical articulatory feature detectors," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico, December 2005.

[20] E. Fosler-Lussier, C. A. Rytting, and S. Srinivasan. "Phonetic ignorance is bliss: Investigating the effects of phonetic information reduction on ASR performance," *Proc. InterSpeech*, Lisbon Portugal, September 2005.

[21] Y. Wang and E. Fosler-Lussier, "Integrating phonetic boundary discrimination explicitly into HMM systems," *Proc. InterSpeech*, Pittsburgh, PA, Sept. 2006.

[22] C. Ma and C.-H. Lee, "Speaker Verification Based on Combining Speaker Individuality Parameter Selection and Decisions," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, San Juan, Dec. 2005.

[23] Q. Fu and B.-H. Juang, "Segment-Based Phonetic Class Detection Using Minimum Verification Error (MVE) Training," *Proc. InterSpeech*, Lisbon, September 2005.

[24] J. Morris and E. Fosler-Lussier, "Combining phonetic attributes using conditional random fields," *Proc. InterSpeech*, Pittsburgh, PA, Sept. 2006.

[25] J. Morris and E. Fosler-Lussier. "Discriminative Phonetic Recognition with Conditional Random Fields," HLT-NAACL Workshop on Computationally Hard Problems and Joint Inference in Speech and Language Proc., 2006.

[26] R. A. Sukkar and C.-H. Lee, "Vocabulary Independent Discriminative Utterance Verification for Non-Keyword Rejection in Subword Based Speech Recognition," *IEEE Trans. on Audio and Speech Proc*, pp. 420-429, 1996.

[27] Y. Tsao, J. Li and C.-H. Lee, "A Study on Separation between Acoustic Models and Its Applications," *Proc. InterSpeech*, Lisbon, Portugal, September 2005.

[28] N. Chomsky and M. Halle, *The Sound Pattern of English*, Harper and Row, 1968.

[29] S. Dusan and L. R. Rabiner, "On Integrating Insights from Human Speech Perception into Automatic Speech Recognition," *Proc. InterSpeech*, Lisbon, Sept. 2005.

[30] S. Dusan, "On the relevance of some spectral and temporal patterns for vowel classification," *Speech Communication*, Vol. 49(1), pp. 71-82, 2007.

[31] S. Dusan and L. Rabiner, "On the Relation between Maximum Spectral Transition Positions and Phone Boundaries," *Proc. InterSpeech*, Pittsburg, PA, Sept. 2006.

[32] S. M. Siniscalchi, J. Li, and C.-H. Lee, "A Study on Lattice Rescoring with Knowledge Scores for Automatic Speech Recognition," *Proc. InterSpeech*, Pittsburgh, Sept. 2006.

[33] C. Ma, Y. Tsao, and C.-H. Lee, "A Study on Detection Based Automatic Speech Recognition," *Proc. InterSpeech*, Pittsburgh, PA, Sept. 2006.