

An Information-Extraction Approach to Speech Analysis and Processing

Chin-Hui Lee

School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Abstract

It is believed that the current capabilities of the state-of-the-art automatic speech recognition (ASR) technologies can be further enhanced by analysis and processing techniques that can take advantage of the full set of acoustic and linguistic information existing at various levels of the speech knowledge hierarchy. This calls for a bottom-up knowledge integration framework that links speech processing with information extraction, by spotting speech cues with a bank of attribute detectors, weighing and combining acoustic evidences to form cognitive hypotheses, and verifying these theories until a consistent recognition decision can be reached. The recently proposed automatic speech attribute transcription paradigm is an attempt to extract relevant speech information, organize these evidences, and integrate them into the already-powerful, top-down hidden Markov modeling framework to facilitate a knowledge-rich, data-driven framework that will lower the entry barriers to ASR research, and still leverage upon many recent advances in automatic model learning. Everyone in and out of the current ASR community will be able to contribute to this worthwhile effort.

Index Terms: automatic speech attribute transcription, automatic speech recognition, speech attribute detection, bottom-up knowledge integration

1. Introduction

At the current stage of a crossroad for the field of automatic speech recognition (ASR) it is instructive to examine some of the key developments in ASR in the past few decades and contemplate on potential new directions that might lead us to a final solution. On one hand it is clear that we have made tremendous technology advances since the ASR community adopted an information-theoretic perspective of channel modeling for encoding speech from language, and formulated the ASR problem as a channel decoding paradigm [1].

This statistical pattern matching approach to ASR is considered as a paradigm shift from the traditional perspective before the 1970s. The ease of learning speech and language models from data therefore triggered a wave of almost four decades of fast technology progresses for ASR based on an integrated pattern modeling and decoding framework, later known as hidden Markov models (HMMs, e.g., [2]), without the need to use detailed knowledge about a target language, sometimes referred to as *knowledge-implicit* or *knowledge-ignorant* modeling [3]. It offers an advantage for automatic model learning via a rigorous mathematical formulation and global optimization by integrating all the required knowledge sources for a task into a finite state network representation, known as *top-down knowledge integration*, ready for DP-based optimal decoding [4, 5].

On the other hand it is also clear that the ASR problem is still far from being solved judging from the limited deployment of spoken language products and services worldwide, the fragile nature of performance robustness for the state-of-the-

art ASR systems, and the slowdown in performance improvement when the amount of training data to learn acoustic and languages model is no longer a major limitation in designing ASR systems. An interesting question to ask is: "Can we revisit many knowledge sources created in speech science, sometimes known as *bottom-up knowledge integration*, and combine it with data-driven learning techniques so that we can develop a *knowledge-rich* framework for modeling speech and language [3] to enhance the current capabilities of ASR technologies?"

It could be argued that the ASR problem is still too difficult that we could not simply rely on a single decoding equation in channel modeling to provide all the solutions. Many speech researchers would agree that the entry barrier to today's competitive ASR research is simply too high for most small groups to be able to make significant impacts. Moreover, we might want to divide up the big and not-easy-to-solve problem into a set of small and manageable sub-problems and conquer them one-by-one. This divide-and-conquer strategy enables a plug-and-play mode for individual researchers and small groups to contribute their best modules to the overall system design via what was called a *Collaborative ASR Community of the 21st Century* [6].

There have been many attempts to find the next paradigm shift. One of them is exploiting the *distinctive features* of speech (e.g., [7, 8]). By organizing such information in a systematic manner, ASR could, in theory, be performed by first identifying and labeling the sequence of *speech attributes*, and then followed by decoding the corresponding sequence of words using lexical access as demonstrated in spectrogram reading by human experts [9]. Using bottom-up knowledge integration [10] human speech recognition (HSR) also performed much better than ASR in most benchmark tests [11].

The speech cues, or attributes, or events, can serve as *acoustic landmarks* [12, 13, 14, 15], or sometimes referred to as "*islands of reliability*" in the ocean of competing speech events, to improve bottom-up knowledge integration. Recently a detection approach to ASR, called automatic speech attribute transcription (ASAT) [16], was proposed in another attempt to address research issues in HSR and spectrogram reading through *bottom-up speech attribute detection* and *stage-by-stage knowledge integration*. We will collectively refer to this set of viewpoints as an "*information extraction*" perspective to extract useful acoustic and linguistic information for the purposes of speech recognition and understanding which serves as the central theme for our discussion in this paper.

2. ASR Technology Challenges

Although we have witnessed quite a number of successful implementations of commercial products and services for many different languages, the ASR technology is still rather fragile. In contrast to the model-based pattern matching approach to *extracting information* from speech, a collection of *signal-based* algorithms needs to be developed in order to detect acoustic

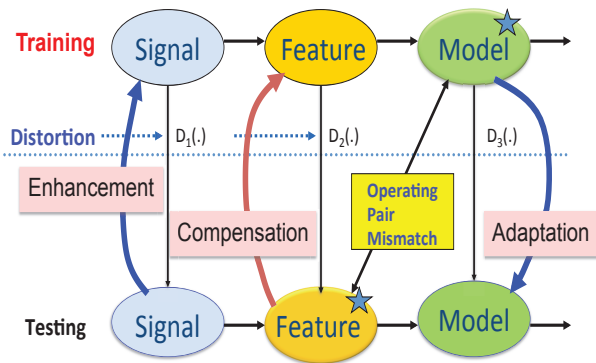


Figure 1: The two starred blocks indicate that the model obtained in training and the features obtained in testing give acoustic mismatches if the testing environments are different from the training conditions, resulting in an operating pair mismatch.

landmarks, such as vowels, glides and fricatives, in adverse conditions. They could serve good purposes for selecting good data segments and designing signal-specific speech enhancement, feature compensation, and model adaptation algorithms for reliable information extraction.

2.1. Challenges in Model Estimation and Robustness

In an evaluation, we tested the WSJ0 task on the Nov92 set, and found that a 4% word error rate (WER) can be achieved with a trigram language model. However, it was degraded to 7% with a bigram language model. More seriously, when every word can follow every word with equal probabilities the WER went as high as 70%. The topic of parameter estimation and adaptation is one of the most fruitful areas in the field of ASR in recent years. Since it is not practical to collect a large set of speech and text examples, spoken and written by a large population over all possible combinations of signal conditions, it is likely a mismatch in training and testing conditions is a major source of errors for pattern matching systems.

The mismatch can be conceptually viewed in the signal, feature, or model space as shown in Figure 1 where a *maximum likelihood stochastic matching* framework was proposed to address the ASR robustness issues caused by this mismatch [17]. The mismatch can be modeled by a distortion in the signal space, D_1 , and handled by speech enhancement. We can also consider a feature space distortion, D_2 , and perform feature compensation. Finally, we can view the mismatched situation in the model space with a transformation D_3 that maps the trained models into the test environment via adaptation [18].

2.2. Attribute Inconsistency in Integrated Search

For example in the WSJ task, we had observed that a conventional system often confused the word "Safra", with the phrase "Stock for". Nonetheless, to recognize the word *stock* it requires the presence of two stop sounds, /t/ and /k/, in the region of a vowel. This can be checked by visually inspecting the spectrogram in the upper panel of Figure 2 which did not show the presence of stops before and after the middle vowel. Moreover the time evolution of the output posterior probabilities for each unit at each frame, known as a posterigram [19] generated by a bank of detectors for manner of articulation displayed in the

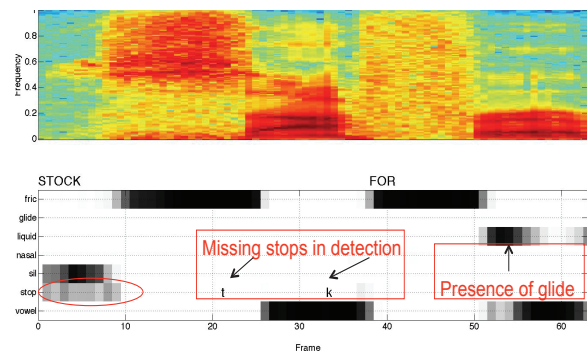


Figure 2: Spectrogram (upper panel) and posterigram (lower panel) for the sentence numbered 446c0210 with focus on the area where the errors occur. A system misrecognized the word "Safra" and generates the transcription "stock for". The time evolution of detecting manner of articulation shows that there are no plosive events in the time span under analysis.

lower panel of Figure 2, clearly indicated that there were no stop events in the area where the mistake occurred, and it also signaled the presence of a glide (/r/ in this case) followed by a vowel at the end of the time-span under analysis. If this information could be properly extracted and included into integrated search, these errors could have been avoided as shown in our lattice rescoring study [20]. Other properties in speech production and auditory perception (e.g., [21]) can also be utilized as attribute information to improve ASR accuracy.

2.3. Prosody Inconsistency in Integrated Search

In a similar way, suprasegmental information, such as prosody, could not be easily cast into the FSN specification when performing top-down knowledge integration. Potential errors could be corrected by using pitch contours and duration features, as demonstrated by a visual inspection of Figure 3¹. In the top panel the waveform for the WSJ sentence *if the Fed pushes the dollar higher, it may curb the demand for U.S. exports* in the time span between 2.72 and 3.32 seconds is displayed.

Specifically, the phrase "it may curb" was misrecognized as "and maker of". The panel below shows the frame energy, whereas the F0 contour is shown in the third panel. The recognized phone and word sequences are reported in the fourth and fifth panels, respectively, and the reference phone and word transcriptions are displayed in the sixth and seventh panels, respectively. Analysis on the second and third plots reveals two inconsistencies in recognizing the middle word "maker" in the phrase, namely: (i) the F0 for the segment "ker" is too high with respect to that for the preceding segment "ma" which puts a strong stressed syllable in the middle of the word, and (ii) the "glottal closure" of 60 milliseconds of the stop sound in "maker" is too long. It should be a "stop gap" of an unvoiced stop as in the correct but misrecognized word "curb" instead.

We have conducted a study on incorporating prosodic models into a state-of-the-art ASR system and managed to significantly reduce the word error rate for the WSJ0 task. We will report this set of new findings in a future paper. Other techniques to improve the extraction of prosody information would likely provide additional gains to ASR performance.

¹Dr. Chen-Yu Chiang of NCTU, Taiwan, had generated this figure.

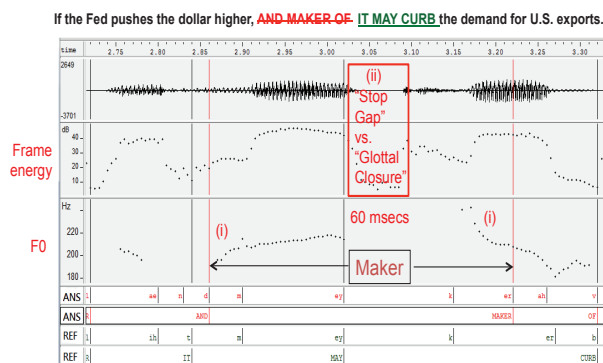


Figure 3: Prosodic analysis shows two inconsistencies for the recognized phrase: (i) the F0 for the segment "ker" is too high with respect to that for the preceding segment "ma", and (ii) "glottal closure" of the stop sound in "maker" is too long.

2.4. Challenges in Spontaneous Speech Processing

Although we have achieved low word error rates in many LVCSR tasks, the high accuracy usually could not be extended to recognizing spontaneous speech. Due to incomplete task specifications in defining most speech recognition and understanding tasks a top-down knowledge integration scenario usually could not maintain the same level of performance obtained in most LVCSR systems. Knowing which part of an utterance to process and to ignore to accomplish a partial understanding of any given utterance is very critical in order to handle spontaneous speech. In [22], it was found that a combined key-phrase recognition and verification strategy worked well especially for handling *ill-formed* utterances. It is designed to detect semantically significant speech segments and reject irrelevant portions.

3. Automatic Speech Attribute Transcription

The speech signal contains a rich set of information that facilitates human auditory perception and communication, beyond a simple word transcription. 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 them speech attributes.

ASR can now be extended to ASAT, a process that promises to be knowledge-rich and capable of incorporating multiple levels of information in the knowledge hierarchy into attribute detection, knowledge integration, and evidence verification, as shown in Figure 4. The top panel illustrates the frontend processing which consists of an ensemble of speech analysis and a bank of detectors. And the bottom panels demonstrate a possible stage-by-stage backend knowledge integration process.

3.1. Frontend Attribute Detection

An event detector converts an input speech signal $x(t)$ into a time series which describes the level of presence of a particular attribute, or event, in the input speech utterance over time. This function can be computed as the a posteriori probability evolving over time, or a likelihood ratio of a target model over a contrast model). The bank of detectors consists of a number of such attribute detectors, each being individually and optimally

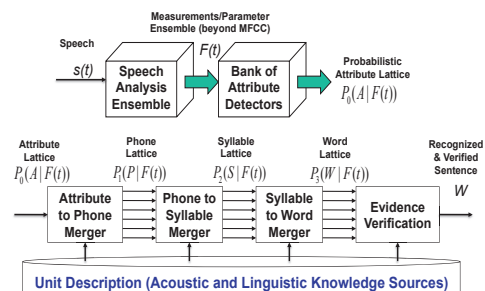


Figure 4: ASAT) – (a) top: speech analysis followed attribute detection to produce an attribute lattice; (b) bottom: stage-by-stage knowledge integration to obtain the recognized sentences.

designed for the detection of a particular event. These attribute properties are often relevant to information to be extracted for 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 detecting voice onset time, and long-term detectors, e.g., for detecting pitch contours, syllables, and word sequences. The conventional frame synchronous analysis, such as MFCC, is thus relaxed in ASAT to accommodate asynchronous detection in which various parameters at different frame rates can be utilized and combined to design attribute-specific event detectors.

Attributes can be used as cues or landmarks [13, 15] in order to identify "islands of reliability" for making local acoustic and linguistic decisions, such as energy concentration and phrase boundaries, without extensive speech modeling. A few clear examples are readily visible in most spectrogram plots, e.g., the vowel and fricative regions in Figure 5.

3.2. Backend Knowledge Integration

Another critical component in the ASAT paradigm is the backend processing. 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 to be fed back for further event merger and knowledge integration.

The procedure produces the evidences needed for a final decision, including a recognized sentence. Each activity function can be modeled by a corresponding neural system. Both activation levels and firing rates have been used in neural encoding and neuron combinations to encode temporal information. Simulating perception of temporal events is of particular interest in auditory perception of speech. The processing of attribute merging can also be interpreted as a form of probabilistic parsing of event lattices. New techniques are needed to accomplish this form of event merging.

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 to improving our understanding of speech, as well as enhancing the ASR performance.

4. ASAT Applications

Next we present several attribute-based applications: lattice rescoring [20], language-universal phone recognition [23], spoken language identification [24], and bottom-up, stage-by-stage LVCSR [25]. We hope to inspire other new applications based on the detector-based frontend processing. Potential new insight that may be gained by attribute detection is discussed first.

4.1. Visible Detection Curves of Information Displays

We now analyze detection score plots more closely in Figures 5. It is interesting to note that the correct transcript can still be read out by following the evolution of the event detection process over time. This outcome is also in line with spectrogram reading by trained experts (e.g., [9]) based on knowledge in acoustic phonetics (e.g., [26]). The detector scores here were normalized between 0 and 1, ranging from an absence of an acoustic property to a fully presence of a speech cue. The value of these detection scores is a good indication of the activity levels for the speech events of interest. Therefore it provides a new visualization tool in addition to the conventional spectrogram. Error analysis has always played a crucial role in providing diagnostic information to improving algorithms in the history of ASR technology development. With the extracted speech information revealed in the new visualization tool, insight may be developed in understanding human speech.

For example we can see sound transition behavior clearly displayed in the region from Segment 6 to Segment 8, going from phone /eh/ to /aa/ with a rising activity from the preceding vowel into the glide sound /l/ in Segment 7 then falling away into the following vowel. We can also observe an overlapping nature of nasalized vowel at the ending of Segment 8 and the beginning of Segment 9. The double stop sound regions in Segments 13 and 14 are also interesting to notice. The large overlapping region for the two-candidate Segment 21 indicates that the glide sound /r/ heavily influence articulation in its surrounding phones with a low level vowel activity showing up between Segments 21 and 22 on the detector plot for the vowel manner.

4.2. Attribute-Based Lattice Rescoring

Attribute-based lattice rescoring was successfully implemented in three different contexts [20]: (i) continuous phone recognition, (ii) domain-independent connected digit recognition with no digit-specific training, and (iii) LVCSR. They all demonstrated improved system performance. Prosody-based lattice rescoring will be reported in a future paper.

4.3. Language-Universal Attribute and Phone Recognition

Fundamental speech attributes, such as voicing, nasality, and frication, could be shared across many different languages, so they could also be used to derive a language-universal set of speech units. It is worth noticing that these phonetic features (attributes) had already been used to identify a common knowledge source in several studies (e.g., [27, 28]). We have studied cross-language and language-universal attribute and phone recognition. More detail can be found in [23].

4.4. Automatic Spoken Language Recognition

A novel universal acoustic characterization approach to spoken language recognition (SLR) under the ASAT framework was proposed in [24]. The key idea was to describe any spoken language with a common set of speech attributes that can be

defined "universally" across all spoken languages. A language can then be recognized using sequences of tokenized speech attributes and their co-occurrence statistics without utilizing the conventional lexical or phonotactic representations.

4.5. ASAT-Based Bottom-Up LVCSR

Our first attempt to implement a bottom-up detection-based LVCSR can be found in [25] using weighted finite state machines (WFSMs) [29]. ASR is accomplished in a bottom-up fashion by performing backend lexical access and syntax knowledge integration over the output of our detection-based frontend, which generates frame-level speech attribute detection scores and phone posterior probabilities. Modular search is made possible by two main factors: (i) high-accuracy detection of acoustic information to generate high-quality lattices at every stage of information processing; and (ii) low-error pruning of the generated lattices to reduce search errors during modular knowledge integration from low-level attributes to high-level recognized sentences.

5. Conclusions

Although the ASR community had made tremendous technology advances in the past, the ASR problem is still far from being solved. We will see continuing development of spoken language products, services and applications to meet the rising demands from the society. In the meantime to make fast progresses again leveraging upon what the technology and resources the speech community had accumulated over the past years this is a good point in time to explore new directions in ASR. In this paper we have presented an information extraction perspective and an ASAT framework that may serve as a paradigm shift for designing next-generation ASR systems. Many preliminary experiments have demonstrated the extra benefits offered by ASAT. To have similar impacts generated by HMM in the last four decades we expect more research is needed to make ASAT a successful vehicle. Everyone inside and outside the current ASR community can contribute to this divide-and-conquer effort.

6. Acknowledgment

The author would like to thank the NSF ITR Program for supporting the ASAT project. He would also acknowledge the contributions of Prof. Sabato Marco Siniscalchi, Kore University of Enna, Sicily, Italy for his co-authorship of an invite paper in a Special Issue of the Proceedings of the IEEE on Speech Information Processing to be published in June 2013. This paper is a concise version of the full paper.

7. References

- [1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, Jul./Oct. 1948.
- [2] L. R. Rabiner, "A tutorial on hidden Markov models and selected application in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [3] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition," in *Proc. Interspeech*, Jeju Island, Korea, Oct. 2004, pp. 109–112.
- [4] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 5, no. 2, pp. 179–190, 1983.

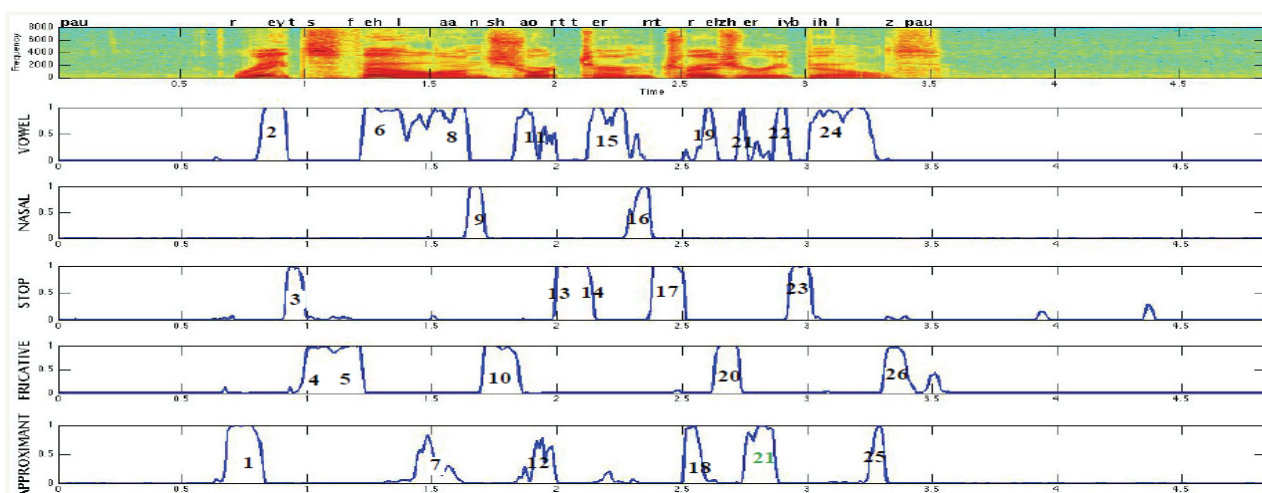


Figure 5: Detection curves of manner of articulation for the sentence numbered 440c20t (*RATES FELL ON SHORT TERM TREASURY BILLS*). The correct transcript can be read out by following the time evolution of the detected attribute events.

- [5] H. Ney and S. Ortmanns, "Progresses in dynamic programming search for LVCSR," *Proc. IEEE*, vol. 88, no. 8, pp. 1224–1240, 2000.
- [6] C.-H. Lee, "On automatic speech recognition at the dawn of the 21st century," *IEICE Trans. on Information and Systems, Special Issue on Speech Information Processing*, vol. 86-D, no. 3, pp. 377–396, 2003.
- [7] G. Fant, *Speech Sounds and Features*. MIT Press, 1973.
- [8] D. H. Klatt, "Review of the ARPA speech understanding project," *J. Acoust. Soc. Am.*, vol. 62, no. 6, pp. 1345–1366, 1977.
- [9] V. W. Zue, "Acoustic-phonetic knowledge representation: implications from spectrograms reading experiments," in *NATO ASI on Speech Recognition*, Bonas, France, 1981.
- [10] J. Allen, "How do humans process and recognize speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, Oct. 1994.
- [11] R. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, 1997.
- [12] V. W. Zue, J. Glass, M. Phillips, and S. Seneff, "The MIT summit speech recognition system: A progress report," in *Proc. DARPA Speech and Natural Language Workshop*, Philadelphia, PA, USA, Feb. 1989, pp. 179–189.
- [13] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," *J. Acoust. Soc. Am.*, vol. 100, no. 5, pp. 3417–3430, 1996.
- [14] A. Juneja and C. Espy-Wilson, "An event-based acoustic-phonetic approach to speech segmentation and E-Set recognition," in *Proc. ICPhS*, Barcelona, Spain, May 2003.
- [15] M. Hasegawa-Johnson, J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan, J. Muller, K. Sonmez, and T. Wang, "Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop," in *Proc. ICASSP*, Philadelphia, PA, USA, Mar. 2005, pp. 213–216.
- [16] C.-H. Lee, M. A. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, B.-H. Juang, and L. R. Rabiner, "An overview on automatic speech attribute transcription (ASAT)," in *Proc. Interspeech*, Antwerp, Belgium, Aug 2007, pp. 1825–1828.
- [17] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 3, pp. 190–202, May 1996.
- [18] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proc. IEEE*, vol. 88, no. 8, pp. 1241–1269, 2000.
- [19] P. Fousek and H. Hermansky, "Towards ASR based on hierarchical posterior-based keyword recognition," in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 433–436.
- [20] S. M. Siniscalchi and C.-H. Lee, "A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition," *Speech Communication*, vol. 51, pp. 1139–1153, 2009.
- [21] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, 2nd ed. Springer-Verlag, 1972.
- [22] T. Kawahara, C.-H. Lee, and B.-H. Juang, "Flexible speech understanding based on combined key-phrase detection and verification," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 6, pp. 558–568, 1998.
- [23] S. M. Siniscalchi, D.-C. Lyu, T. Svendsen, and C.-H. Lee, "Experiments on cross-language attribute detection and phone recognition with minimal target specific training data," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 3, pp. 875–887, Mar. 2012.
- [24] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Computer Speech & Language*, 2012, to be published.
- [25] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "A bottom-up step-wise knowledge-integration approach to large vocabulary continuous speech recognition using weighted finite state machines," in *Proc. Interspeech*, Florence, Italy, Sept. 2011, pp. 901–904.
- [26] K. Stevens, *Acoustic Phonetics*. MIT Press, 1998.
- [27] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, 2001.
- [28] H. Lin, L. Deng, D. Yu, Y.-F. Gong, A. Acero, and C.-H. Lee, "A study on multilingual acoustic modeling for large vocabulary ASR," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 4333–4336.
- [29] M. Mohri, "Finite-state transducers in language and speech processing," *Computational Linguistics*, vol. 23, no. 2, pp. 269–311, 1997.