LOQUENS 1(1)
January 2014, e005
eISSN 2386-2637
doi: http://dx.doi.org/10.3989/loquens.2014.005

An attribute detection based approach to automatic speech processing

Sabato Marco Siniscalchi[1] and Chin-Hui Lee[2]

[1] University of Enna "Kore", [2] Georgia Institute of Technology
e-mail: marco.siniscalchi@unikore.it, chl@ece.gatech.edu

Citation / Cómo citar este artículo: Siniscalchi, S. M., and Lee, C.-H. (2014). An attribute detection based approach to automatic speech processing. *Loquens, 1*(1), e005. doi: http://dx.doi.org/10.3989/loquens.2014.005

ABSTRACT: State-of-the-art automatic speech and speaker recognition systems are often built with a pattern matching framework that has proven to achieve low recognition error rates for a variety of resource-rich tasks when the volume of speech and text examples to build statistical acoustic and language models is plentiful, and the speaker, acoustics and language conditions follow a rigid protocol. However, because of the "blackbox" top-down knowledge integration approach, such systems cannot easily leverage a rich set of knowledge sources already available in the literature on speech, acoustics and languages. In this paper, we present a bottom-up approach to knowledge integration, called automatic speech attribute transcription (ASAT), which is intended to be "knowledge-rich", so that new and existing knowledge sources can be verified and integrated into current spoken language systems to improve recognition accuracy and system robustness. Since the ASAT framework offers a "divide-and-conquer" strategy and a "plug-and-play" game plan, it will facilitate a *cooperative speech processing community* that every researcher can contribute to, with a view to improving speech processing capabilities which are currently not easily accessible to researchers in the speech science community.

KEYWORDS: speech attribute detection; knowledge-rich systems; artificial neural networks; hidden Markov models

RESUMEN: *Una estrategia de procesamiento automático del habla basada en la detección de atributos.*- Los sistemas más novedosos de reconocimiento automático de habla y de locutor suelen basarse en un sistema de coincidencia de patrones. Gracias a este modo de trabajo, se han obtenido unos bajos índices de error de reconocimiento para una variedad de tareas ricas en recursos, cuando se aporta una cantidad abundante de ejemplos de habla y texto para el entrenamiento estadístico de los modelos acústicos y de lenguaje, y siempre que el locutor y las condiciones acústicas y lingüísticas sigan un protocolo estricto. Sin embargo, debido a su aplicación de un proceso ciego de integración del conocimiento de arriba a abajo, dichos sistemas no pueden aprovechar fácilmente toda una serie de conocimientos ya disponibles en la literatura sobre el habla, la acústica y las lenguas. En este artículo presentamos una aproximación de abajo a arriba a la integración del conocimiento, llamada transcripción automática de atributos del habla (conocida en inglés como *automatic speech attribute transcription*, ASAT). Dicho enfoque pretende ser "rico en conocimiento", con el fin de poder verificar las fuentes de conocimiento, tanto nuevas como ya existentes, e integrarlas en los actuales sistemas de lengua hablada para mejorar la precisión del reconocimiento y la robustez del sistema. Dado que ASAT ofrece una estrategia de tipo "divide y vencerás" y un plan de juego de "instalación y uso inmediato" (en inglés, *plug-and-play*), esto facilitará una *comunidad cooperativa de procesamiento del habla* a la que todo investigador pueda contribuir con vistas a mejorar la capacidad de procesamiento del habla, que en la actualidad no es fácilmente accesible a los investigadores de la comunidad de las ciencias del habla.

PALABRAS CLAVE: detección de los atributos del habla; sistemas ricos en conocimientos; redes neuronales artificiales; modelos ocultos de Markov

1. INTRODUCTION

Automatic speech recognition (ASR) is commonly addressed using data-driven approaches and a number of fine textbooks and reference books have been published on this topic (De Mori, 1998; Jelinek, 1997; Lee, Soong, & Paliwal, 1996; O'Shaughnessy, 2000; Rabiner & Juang, 1993). The key features of the conventional approach are the use of a decoding strategy based on dynamic programming (DP; e.g., Bellman, 1957; Bellman & Dreyfus, 1962; Ney & Ortmanns, 2000), along with a Markov inference framework (Baker, 1975; Baum, 1972; Baum & Petrie, 1966; Baum, Petrie, Soules, & Weiss, 1970). The ease of learning speech and language models from data has triggered, in the last 40 years, a wave of progress using fast technology for ASR based on this integrated pattern modeling and decoding framework, later known as hidden Markov models (HMMs; e.g., Lee & Huo, 2000; Rabiner, 1989).

However, in recent years, technological progress has slowed down significantly and most research groups are searching for the next big wave to move ASR forward. The fragile nature of ASR system design requires new technological breakthroughs before conversation-based systems really become a ubiquitous user interface mode, able to compete with conventional graphical user interfaces using a "point-'n'-click" device like a mouse, or touch-sensitive screens.

It could be argued that the ASR problem is still too difficult and that the community could not simply rely on a single ASR decoding equation in the Shannon channel modeling paradigm (Shannon, 1948) to provide all the answers. Many speech researchers would agree that the entry barrier to competitive ASR research today is simply too high for most speech groups to be able to make any significant impacts.

Moreover, we might want to divide up the big and not-easy-to-solve ASR problem into a set of small and manageable issues and conquer them one by one. The "divide-and-conquer" strategy enables a "plug-and-play" mode for small individual researchers to contribute their best modules to the overall system through what was called a "collaborative ASR community of the 21st century" (Lee, 2003).

It seems that the missing link between human speech recognition (HSR) and ASR lies in designing a bank of "perfect" feature detectors to serve as "cues" for further processing. The automatic speech attribute transcription (ASAT; Lee et al., 2007; Lee & Siniscalchi, 2013) framework recently proposed is an attempt to bridge this gap and address research issues in HSR. The key idea is to design *bottom-up speech attribute detection* followed by a *stage-by-stage knowledge integration* paradigm. 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. It also facilitates a divide-and-conquer strategy so that researchers from different corners of the world can collaborate by contributing their best detectors or knowledge integration modules to plug-and-play in the overall system design.

The rest of the paper is organized as follows. In Section 2 we briefly review the automatic speech attribute transcription (ASAT) framework. In Section 3 we highlight new ASAT capabilities to address the technological limitations of the conventional ASR approach and demonstrate how the divide-and-conquer ASAT approach can overcome some of these limitations. In Section 4 we describe possible future work opportunities within the ASAT framework by leveraging a plug-and-play collaboration plan. Finally we summarize our findings.

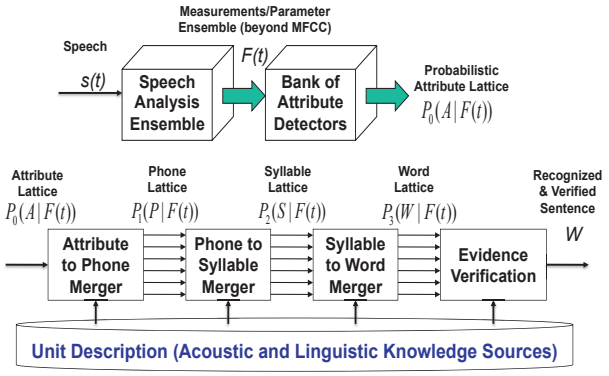
2. AUTOMATIC SPEECH ATTRIBUTE TRANSCRIPTION

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 the ASR and HSR systems, the narrow notion of speech-to-text in ASR has to be expanded to incorporate all the 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. 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.

Based on the above-mentioned 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 (Lee & Siniscalchi, 2013). ASAT therefore promises to be knowledge-rich and capable of combining multiple levels of information in the knowledge hierarchy into attribute detection, evidence verification and integration, as shown in Figure 1. The top panel illustrates the front-end processing which consists of an ensemble of speech analysis and parametrization modules. And the bottom panel demonstrates a possible stage-by-stage back-end knowledge integration process. These two key system components will be described in more detail below. 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 a tight coupling of interdisciplinary studies of speech science and process-

ing. Such a paradigm would eventually lower the entry barrier to competitive ASR research, since speech groups can specialize in a specific block displayed in Figure 1 and yet provide a significant impact.

Figure 1: Automatic speech attribute transcription (ASAT). Top panel: speech analysis ensemble followed by a bank of attribute detectors to produce an attribute lattice. Bottom panel: stage-by-stage knowledge integration from speech attributes to recognized sentences.



In the following paragraphs, the ASAT frontend and backend will be briefly reviewed. The interested reader can refer to Lee and Siniscalchi (2013) for additional details.

2.1. Front-end attribute detection

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 of an attribute, or event, 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 to 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 event.

2.2. Back-end knowledge integration

Another critical component in the ASAT paradigm is the back-end 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. This iterative information fusion process always uses the original event activity functions as the raw cues. A terminating strategy can be instituted by utilizing all the supported attributes.

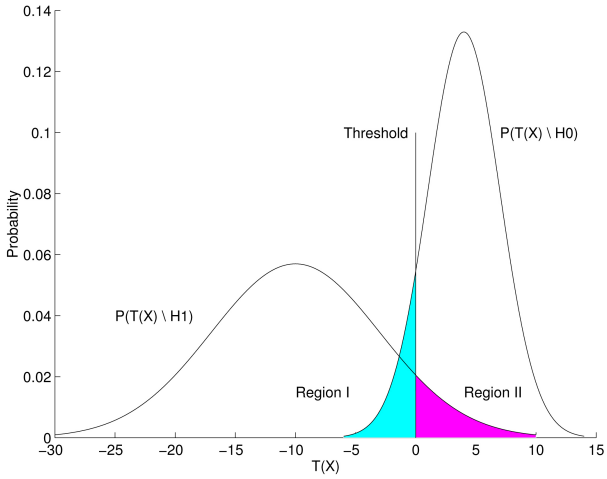
To make use of the features detected, we must combine them in a way that enables us to produce word hypotheses. In essence, this boils down to three problems: (i) combining multiple estimates of the same event to build a stronger hypothesis; (ii) combining estimates of different events to form a new, higher level event with similar time boundaries; and (iii) combining estimates of events sequentially to form longer-term hypotheses. Note that these problems are relatively independent of the level of modeling: while the canonical bottom-up processing sequence would be to combine multiple estimates of each feature, and then combine the features into phones and then words (and word sequences), we envision a highly parallel paradigm that is flexible enough, for example, to combine a feature-based phone detector with a directly-estimated phone detector. In principle, a 20K-word ASR system can be realized with a set of 20,000 single-keyword detectors (Ma & Lee, 2007).

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 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 ASR accuracy. As an example, we found that "knowledge scores" computed with detectors for manner and place of articulation offered a collection of complementary information that can be combined with HMM frame likelihoods to reduce phone and word errors in rescoring (Li, Tsao, & Lee, 2005; Siniscalchi & Lee, 2009; Siniscalchi, Li, & Lee, 2006).

2.3. Event and evidence verification

Verification of patterns is often formulated as a statistical hypothesis testing problem (Lehmann, 1959) as follows: given a test pattern X, we want to test the *null hypothesis*, H_0, against the *alternative hypothesis*, H_1, where H_0 assumes that X is generated from the source, S_0, and H_1 assumes that X is generated from another source, S_1. Event verification, a critical ASAT component, can be formulated in a similar way. There are plenty of techniques available in literature in designing optimal tests if $P(X|H_0)$ and $P(X|H_1)$ are known exactly. However, for most practical verification problems we face in real-world speech and language modeling, we use a set of training examples to estimate the parameters of the distributions of the null and alternative hypotheses. The two competing hypotheses and their overlap indicating the two error types are illustrated in Figure 2. The use of generalized log likelihood ratio (GLLR) was recently proposed as a way to measure separation between competing hypotheses (Tsao, Li, & Lee, 2005). GLLR plots are similar to what is shown in Figure 2 for measuring separation between speech events.

Figure 2: Separation and overlap between competing events with indication of the two error types.



Two types of errors thus exist. The type I error, or false rejection or miss detection rate, is shown in the blue area of Region I in Figure 2, and the type II error, or false acceptance rate, is shown in the magenta area of Region II in Figure 2. The verification performance is often evaluated as a combination of Type I and Type II errors. The related topic of *confidence measure* (CM) has also been intensively studied by many researchers recently. This is due to an increasing number of dialogue applications being developed and deployed in the past few years. In order to have intelligent or human-like interactions in these dialogue applications, it is important to attach to each event a number value that indicates how confident the ASR system is about accepting the recognized event. This number, often referred to as a CM, serves as a reference guide for the dialogue system to provide an appropriate response to its users, just as an intelligent human being is expected to do when interacting with other people. This also demonstrates a clear advantage in HSR: human beings build up their recognition results bottom-up using some forms of confidence measure of events in both the acoustic and linguistic domains.

A few word-level confidence measures have been studied in Kawahara, Lee, & Juang (1998) to improve the rejection of out-of-grammar and out-of-task speech segments for ill-formed utterances often encountered in spontaneous speech. For short and confusable events, such as phones and place of articulation, more research in CM is needed (Sukkar & Lee, 1996). However for some of the events, such as manner of articulation, they exhibit the characteristics of "islands of reliability" with high confidence levels, especially in clean conditions and in human spectrogram reading experiments.

To compare verifiers and detectors, we can plot the receiver operating characteristic (ROC) curves. The areas of Regions I and II in Figure 2 can be estimated, depending on the thresholds used in performing the verification tests. An example is illustrated in Figure 3, in which two systems are compared for detecting the phone /n/ in the first panel. Three operating points for each ROC curve are also indicated. It is clear that the detector producing the left ROC curve performs better than the detector on the right. The ROC curves for detecting the three phones, /w/, /ah /, /n/, respectively, and the word "one", are shown in the second panel of Figure 3, using the corresponding Gaussian density pairs. It can be clearly seen that a 3-phone word "one" is better detected than any of the single phones in the word.

2.4. Attribute detector design

Attribute detectors should be stochastic in nature and designed with data-driven modeling techniques. The goal of each detector is to analyze a speech segment and produce a confidence score or a posterior probability that pertains to some acoustic-phonetic attribute. Generally speaking, both frame- and segment-based data-driven techniques can be used for speech event detection. Frame-based detectors can be realized in several ways, e.g., with artificial neural networks (ANNs), Gaussian mixture models (GMMs), support vector machines (SVM), etc. One of the advantages with ANN-based detectors is that the output scores can simulate the posterior probabilities of an attribute, given the speech signal. On the other hand, segment-based detectors are more reliable in spotting segments of speech (Li & Lee, 2005).

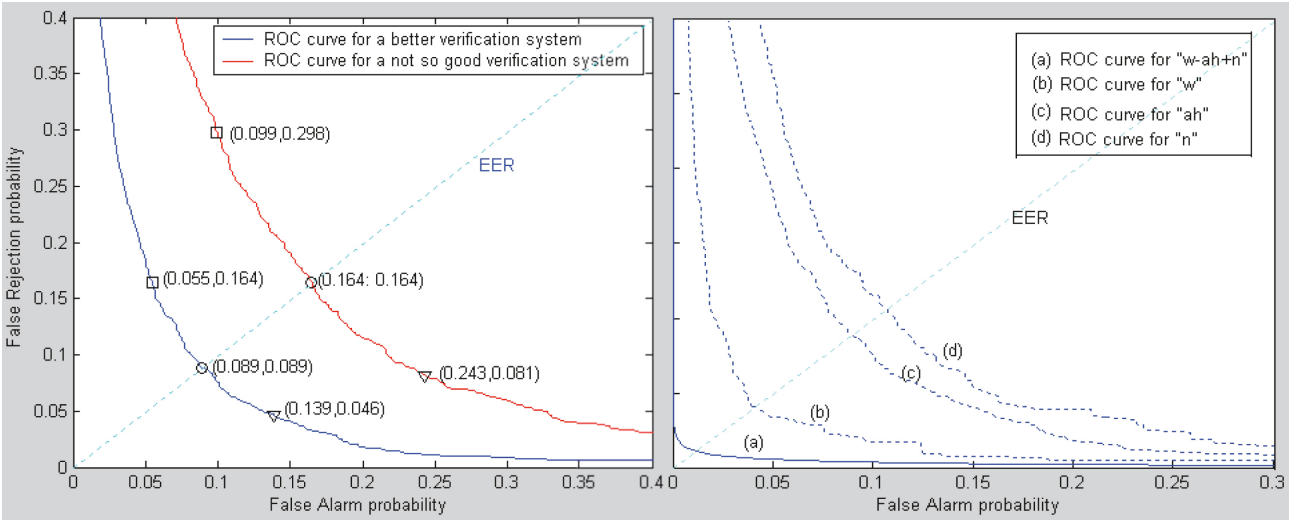
Segment-based detectors can be built by combining frame-based detectors, or with segment models, such as HMMs, which have already been shown to be effective for ASR (Rabiner, 1989). Time-delay neural networks (TDNN) were also shown to be effective in designing segment-based attribute classifiers (Hou, Rabiner, & Dusan, 2007). The reader is referred to a recent PhD thesis (Hou, 2009) detailing the process of building highly accurate TDNN-based classifiers for all the attribute features of interest.

2.5. Back-end merger design

There exist several methods to generate evidence at a sub-word level from articulatory events. For example, starting with manner and place of articulation, a product lattice of degree two may be generated, and a "constrained" search may be performed over this lattice to generate phone level information (Hacioglu, Pellom, & Ward, 2004). Conditional random fields were used in Morris and Folser-Lussier (2006) to generate phone sequences by combining articulatory features.

In our framework, all of the detector outputs are combined with a non-linear function that maps all of the scores between zero and one and at the same time generates phone level information. The non-linear function is realized by a feed-forward multilayer perceptron (MLP) which has a single hidden layer. In more

Figure 3: ROC curves to compare detectors. (Upper) ROC curves comparing two /n/ detectors. (Lower) ROC curves detecting the word and the three phones in "one".



recent work, we have demonstrated that phone accuracies can be boosted using a deep neural network (Deng & Yu, 2011; Mohamed, Dahl, & Hinton, 2009; Seide, Li, & Yu, 2011), as shown in Yu, Siniscalchi, Deng, and Lee (2012). By merging the attribute detector outputs and feeding them into the ANN-based attribute-to-phone mapping merger, we can produce frame-based posterior probabilities, one for each phone of interest, and form a posterior probability feature vector, one for each form being processed. These posteriors can be used as building blocks of language-universal units (e.g., Siniscalchi, Lyu, Svendsen, & Lee, 2012).

3. NEW ASAT PROCESSING CAPABILITIES

With the proposed ASAT paradigm, a set of new speech processing capabilities not easily available or accessible in the state-of-the-art, top-down knowledge integration paradigm can now be explored in the following sections.

An example of visible speech analysis that goes beyond the conventional spectrogram plots is given out in section 3.1. A bottom-up large vocabulary continuous speech recognition (LVCSR) system is described in Section 3.2. Acoustic-phonetic, articulatory, suprasegmental and long-span linguistic constraints can be imposed at different stages of the bottom-up processing to reduce the possibility of inconsistent brute-force decoding results and improve recognition accuracy as demonstrated in Section 3.3.

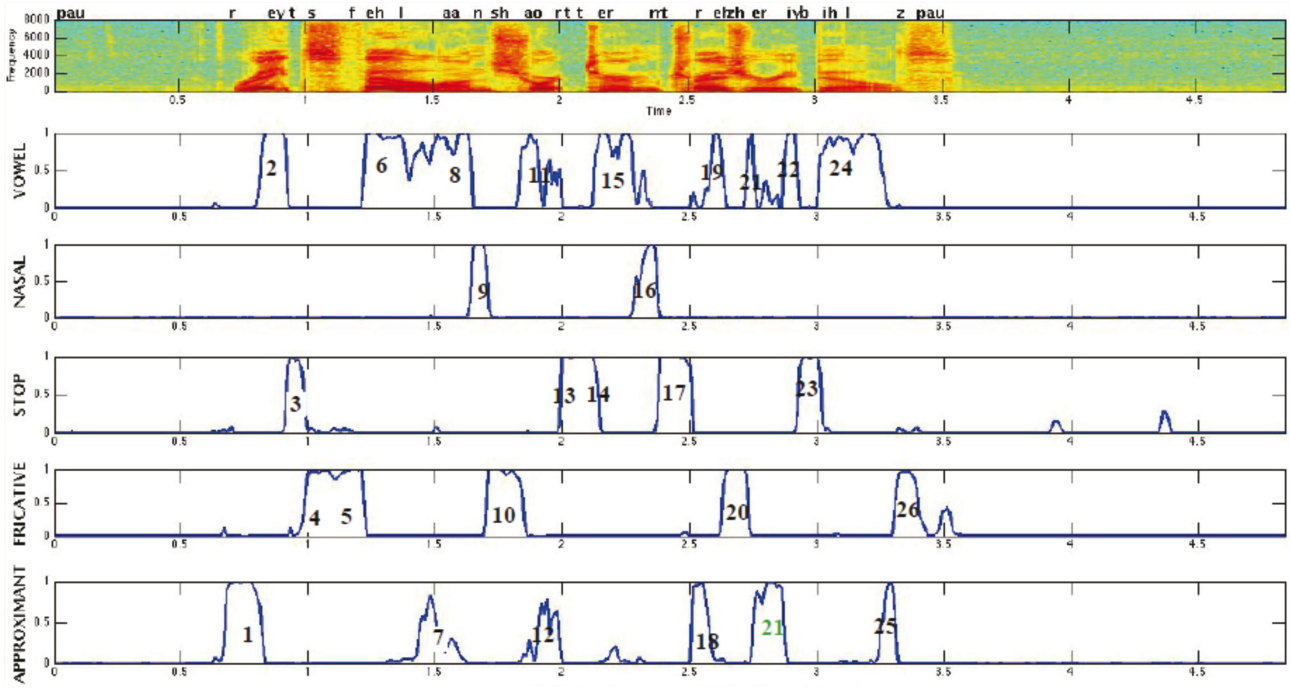
As for modeling of fundamental speech units, we can now explore units that are clearly defined according to acoustic descriptions, such as nasality and frication, and articulatory description, such as back tongue position or dental sounds. These units or cues are now characterized with statistical models and trained with speech examples. However, if they can be detected with signal processing techniques (Rabiner & Schafer, 2011), an enhanced robustness can be expected. The use of speech attributes, such as manner and place of articulation, to serve as language-universal speech units is discussed in Section 3.4. Spoken language recognition (SLR) can now be accomplished by a set of such units to act as tokenizers, as shown in attribute-based SLR (Siniscalchi, Reed, Svendsen, & Lee, 2013). Cross-lingual properties (Siniscalchi et al., 2012) can also be utilized to model speech units in one language and apply them to other languages.

As for speech analysis and feature extraction, the time-synchronous process is a dominant practice: for example, short-time Fourier analysis is often performed at every 10-20 msec on a windowed speech segment of 30-45 msec. This is required in the current decoding paradigm in order to compare likelihood scores on a frame-by-frame basis. However, some speech features, such as pitch and duration, require a longer time frame to process, while features such as voice onset time (VOT) often require a shorter time to process. Within the ASAT framework, asynchronous speech analysis is a key advantage. This will be discussed in Section 3.5. Biologically-inspired and physiologically- motivated speech features can also be explored. Different speech parameters can thus be extracted to design various speech attribute detectors. This important concept will also be illustrated. Pitch and duration information, which is hard to integrate into state-of-the-art speech processing systems, will be shown to improve speech recognition accuracy in Section 3.6.

3.1. Visible speech analysis through attribute detection

We now analyze detection score plots more closely. Figure 4 displays a longer sentence. It is interesting to notice that the correct transcript can still be read out by following the evolution of the event detection process

Figure 4: Detection curves of manner of articulation for the sentence numbered 440c20t (*RATES FELL ON SHORT TERM TREASURY BILLS*) of the SI-84 data set (Paul & Baker, 1992). The correct transcript can still be read out by following the time evolution of detection of the attribute events.



over time. This outcome is also in line with spectrogram reading by trained experts based on knowledge in acoustic-phonetics (e.g. Zue, 1981). The detector scores here were normalized between 0 and 1, ranging from an absence of an acoustic property to the full 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. It therefore provides a new visualization tool in addition to the conventional spectrogram plot shown in the top panel of Figure 4.

Error analysis has always played a crucial role in providing diagnostic information to improve ASR algorithms in the history of ASR technology development. With the extracted speech cue information revealed in the new visualization tool, insight can also be developed in understanding human speech. It also provides a good tool for educating a new generation of speech students.

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 rising activity from the preceding vowel into the approximant 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 approximant sound /ɹ/ heavily influences articulation in the surrounding phones, with a low level vowel activity showing up between Segments 21 and 22 on the detector plot for vowel manner.

It is clear the detector score plots displayed in Figure 4 provide a rich set of information not commonly available to researchers that are not expert-trained in spectrogram reading. It also reinforces additional advantages we intend to exploit in the information-extraction perspective we have highlighted throughout this paper.

3.2. Bottom-up LVCSR

The conventional top-down integrated decoding framework often hampers the definition of generic knowledge sources (e.g., Gauvain & Lamel, 2000; Ney & Ortmanns, 2000) that can be used in different domains. Therefore applications for a new knowledge domain need to be built almost from scratch. Furthermore, the effectiveness of the integrated search diminishes when dealing with unconstrained and possibly ill-formed speech inputs, since more complex language models are needed for handling spontaneous speech phenomena along with much larger lexicons. On the other hand, in the *modular* approach (Siniscalchi et al., 2013) the recognized sentence can be obtained by performing unit matching, lexical matching, and syntactic and semantic analysis in a stage-by-stage, sequential manner. As long as the interface between the adjacent decoding modules can be completely specified, each module can be designed and tested separately.

Our first attempt to implement a bottom-up detection-based LVCSR (Siniscalchi, Svendsen, & Lee, 2011) using weighted finite-state machines (WFSMs; Mohri,

1997) is now presented. ASR is accomplished in a bottom-up fashion by performing back-end 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. Decoupled recognition is made possible by two main factors: (i) high-accuracy detection of acoustic information in order to generate high-quality lattices at every stage of the acoustic and linguistic information processing; and (ii) low-error pruning of the generated lattices in order to reduce search errors likely to occur when trying to minimize the possibility of memory overflow in using the AT&T WFSM tool.

Figure 5: Decoupled, bottom-up, detection based LVCSR with a bigram language model (LM).

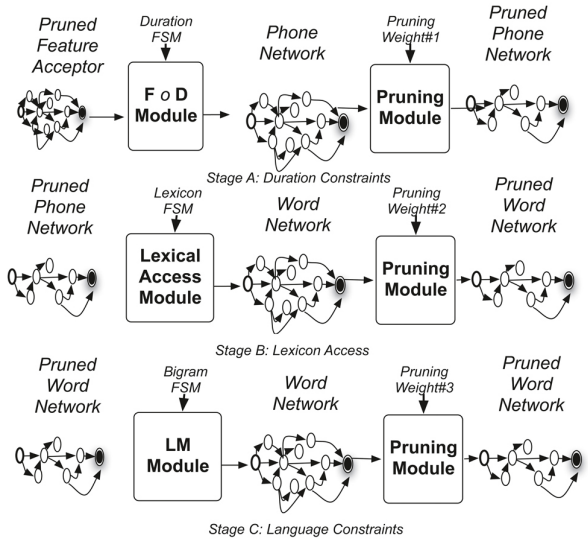


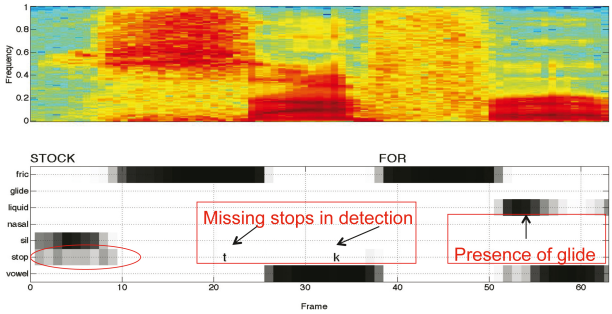
Figure 5 shows how the decoupled approach is implemented in practice using WFSMs. The original feature acceptor F is pruned and composed with the duration transducer D to generate a recognition network (lattice) at the phone-level *phoneRN*. Therefore, the next combination is *phoneRN* o L, which gives a word level recognition network by integrating lexical knowledge. After pruning, this word-level network is composed with G_{bigram} to integrate bigram language model (LM) information. The final grammar-constrained word-level *dRN* is thus generated and sent to the trigram LM rescoring module to re-order the decoding paths embedded in *dRN*. The output of this step is a word-level lattice over which either the best path or the N-best list is computed and delivered.

3.3. Knowledge-based constraints to limit nonsense recognition results

Let us consider inconsistency with acoustic-phonetic evidence in the integrated search. In the Wall Street Journal (WSJ) task (Paul & Baker, 1992), we had observed that a conventional ASR system often confused the word "Safra", with the phrase "Stock for". Nonetheless, to recognize the word *stock* it requires the presence of two stop phones, /t/ and /k/, in the region of a vowel. This can be checked by visually inspecting the spectrogram in the upper panel of Figure 6, which did not show the presence of stop sounds before and after the middle vowel.

Moreover the time evolution of the output posterior probabilities for each unit in each frame, known as a posteriogram (Fousek & Hermansky, 2006), generated by a bank of ANN-based detectors for manner of articulation, displayed in the lower panel of Figure 6, clearly indicated that there were no stop events in the area where the mistake occurred, and it also signaled the presence of an approximant (/ɹ/ 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 in the integrated search, these errors could have been avoided.

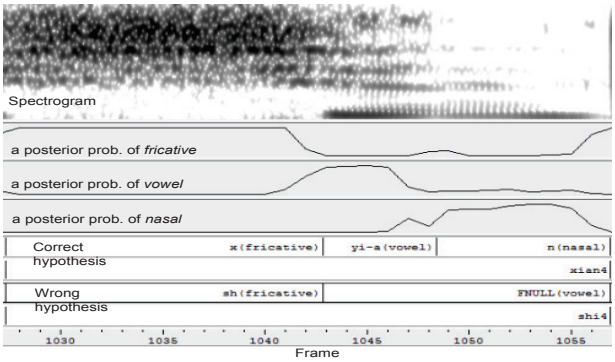
Figure 6: Spectrogram (upper panel) and posteriogram (lower panel) for the sentence numbered 446c0210 of the Nov92 test set with focus on the area of the errors occurring. A conventional LVCSR system missrecognizes the word "Safra," and generates the transcription "Stock for." In the second panel, the time evolution of the posterior probabilities, i.e. the posteriogram, of manner of articulation shows that there are no plosive events in the time span under analysis. Furthermore, wrong word recognition is delivered, although correct manner or articulation detection can be performed.



Another example is given in Figure 7 to show the effectiveness of the cross-language attribute detector. The correct word sequence excerpted from part of the utterance numbered $NCKU_{}060602_0$ in the TCC300 test (Chiang, Siniscalchi, Wang, Chen, & Lee, 2012) set is: "(超級, chao1-ji2, super) (大, da4, large) (縣, xian4, county) (臺北縣, tai2-bei3-xian4, the Taipei County) (其, qi2, its) (縣長, xian4zhang3, County Magistrate) (寶座, bao3-zuo4, post)". The baseline system generated an erroneous word sequence of "(及, ji2, and) (市長, shi4-zhang3, Major)". When applying the attribute scores in rescoring, the word "(市長, shi4-zhang3, Major)" can be corrected. Figure 7 displays the posterior probabilities of the related attributes for the base-syllable "xian4", which is misrecognized as "shi4". Note that the attribute sequence for "xian4" is fricative-vowel-

nasal while that for "shi4" is fricative-vowel. It can be seen that the posterior probabilities of the vowel at the end of the syllable were much lower than those of the nasal attribute. Therefore, the wrong base-syllable hypothesis of "shi4" with no nasal sound was penalized by the syllable structure verifier in favor of the base-syllable of "xian4" with a nasal sound (Chiang et al., 2012).

Figure 7: Recognition results with and without constraints. This example is part of the study presented in Chiang, Siniscalchi, Wang, Chen, and Lee (2012).

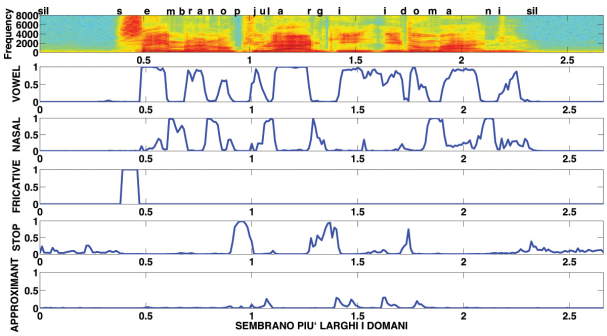


The ASAT detector frontend has been used successfully as a key component for several speech applications, namely: lattice rescoring (Siniscalchi & Lee, 2009; Siniscalchi, Svendsen, & Lee, 2009), language-universal phone recognition (Siniscalchi et al., 2012) and spoken language identification systems (Siniscalchi et al., 2013). In this paper, we review the ASAT automatic spoken language recognition (LRE) system and show how better detectors allow for better system performance. The reader is referred to Lee and Siniscalchi (2013) for a review of all the other ASAT applications.

3.4. Defining and modeling of language-universal acoustic units

Designing good ASR systems with little or no language-specific speech data for resource-limited languages is a challenging research topic. As a consequence, there has been increasing interest in exploring knowledge sharing among a large number of languages, so that a universal set of acoustic phone units can be defined to work for multiple or even for all languages. In Siniscalchi et al. (2012), we have shown that ASAT can play a key role in designing language-universal acoustic models by sharing speech units among all target languages at the acoustic phonetic attribute level. Indeed, it was shown that good cross-language attribute detection and continuous phone recognition performance can be accomplished for "unseen" languages using minimal training data from the target languages to be recognized.

Figure 8: Language-universal attribute detection of an Italian sentence. This figure is taken from Lee and Siniscalchi (2013).



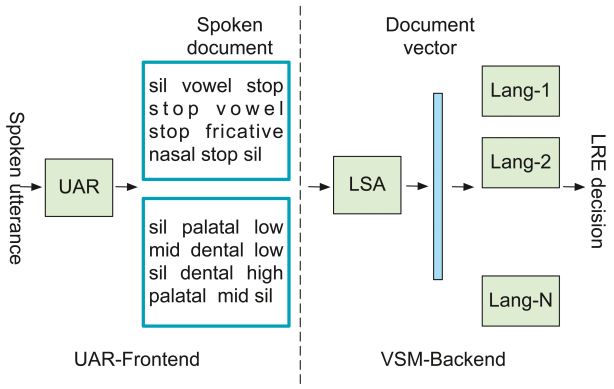
One of the key research challenges that is yet to be addressed in ASR is to build a language universal ASR system for all spoken languages. One way to reach the ultimate goal is to first design language-universal attribute detectors that work for all attributes of interest. In Figure 8, place of articulation detectors trained on the read-English WSJ0 speech data are tested on an Italian utterance from a completely unknown corpus and the detection result is displayed. It is interesting to note that most of the speech attributes were detected with considerable accuracy, except for the combination of /b/ and /r/ at the beginning of the utterance because of the high speed of delivery.

The idea of having a set of universal acoustic units that can be sharable across languages was further exploited in the spoken language recognition (SLR) context. Our aim was to describe a spoken language with a set of speech attributes that can be defined "universally" across all spoken languages (Siniscalchi et al., 2013). The set of universal attributes used in our investigation was defined using manner and place of articulation classes. The silence unit was also taken into account to indicate the absence of an articulation activity. A vector space modeling (VSM) approach to SLR was adopted to accomplish the recognition task, where a spoken utterance is first decoded into a sequence of attributes, independently of its language. The SLR system used in our experiments is shown in Figure 9. It consists of two main blocks: a frontend, shown in the left-hand panel, and a backend, shown in the right-hand panel. The frontend implemented a universal attribute recognizer (UAR) that decodes a spoken utterance into two parallel sequences of manner and place attributes, which have the useful property of being sharable across all spoken languages. The string of attribute symbols mapped spoken utterances into spoken documents. The backend delivered the final SLR decision through a VSM approach in two steps. First, a vector representation of the spoken document is obtained using latent semantic analysis (LSA) (Bellegarda, 2000).

In our SLR studies, the key intuition was that a bag-of-attributes model could universally characterize any spoken language. Furthermore, we observed that error rates decrease with improvements in the attribute reso-

lution of the proposed system. We believe that better results can be attained by designing ad-hoc features for each speech attribute, and expert knowledge could play a critical role here.

Figure 9: Block diagram of the SLR system with UAR-frontend and VSM-backend.

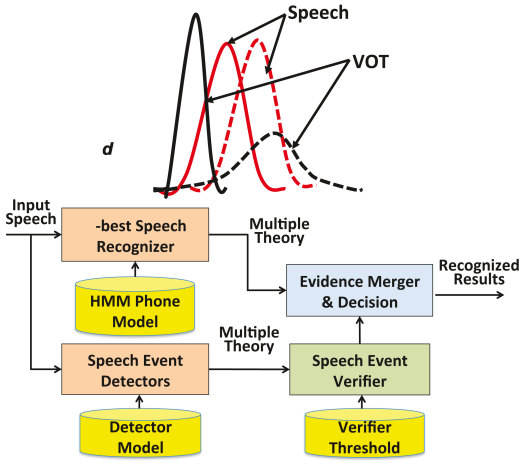


3.5. Asynchronous speech analysis and detector design

Biologically-inspired and perception- motivated signal analysis are considered as promising parameter extraction directions (Jeon & Juang, 2007; Shamma, 2001) because the ASAT paradigm supports parameter extraction at different frame rates for designing a range of speech attribute detectors. Once a collection of speech parameters, *F(t)*, are obtained, they can be used to perform attribute detection which is a critical component in the ASAT paradigm, as shown in the upper panel of Figure 1. Attributes can be used as cues or landmarks in speech (Hasegawa-Johnson et al., 2005; Liu, 1996) in order to identify the "islands of reliability" for making local acoustic and linguistic decisions, such as energy concentration regions and phrase boundaries, without extensive speech modeling.

An attribute detection example was demonstrated in Niyogi, Mitra, and Sondhi (1998) to discriminate voiced and unvoiced stops using voice onset time (VOT) for two-pass English letter recognition as shown in the lower half of Figure 10. In the first stage, a conventional recognizer was used to produce a list of multiple candidates. To further discriminate some of the minimum pairs, such as the English /d/ and /t/, a VOT-based detector (Niyogi, Mitra, & Sondhi, 1998), can be used in the second stage to provide a detailed discrimination as illustrated in the upper half of Figure 10, where we plot a pair of competing distributions of the likelihood ratio for /d/ and /t/ based on 39 cepstral parameters and for a single VOT, respectively. Clearly, the VOT temporal feature produces a pair of curves with better discrimination (i.e., with more separation between them) than those obtained with spectral features alone. By reordering candidates according to VOT, the two stage recognizer gave an error rate of 50% less than that obtained in a state-of-the-art ASR system (Ramesh & Niyogi, 1998). The same notion has been applied to natural number recognition in which resolving the "teen-ty" confusion using a high performance nasal detector is critical.

Figure 10: (Upper) A single VOT parameter is better than 39 cepstral features in discriminating voiced /d/ against unvoiced stop /t/; (Lower) A two-stage speech recognizer incorporating sound-specific event detectors.



3.6. Integration of suprasegmental information

The proposed knowledge-assisted ASR was evaluated on a large Mandarin read speech corpus TCC300 (Association for Computational Linguistics and Chinese Language Processing [ACLCLP], 2013a). The acoustic feature vector used here consists of 38 components (12 MFCC parameters, their first and second order time derivatives, one delta energy and one delta-delta energy) analyzed at a 10-msec frame rate with a 30-msec Hamming window. HMM parameters of the 411 8-state syllables were estimated with maximum mutual information (MMI; Bahl, Brown, de Souza, & Mercer, 1986) using part of the TCC300 training data (274 speakers, about 23 hours). A test set was formed selecting utterances from 19 speakers (about 2 hours) among the TCC300 test data. All test data were paragraph utterances with an average length of 32 seconds.

The SRILM (Stolcke, 2002) toolkit was used to train the factored LM with several text corpora, including (i) Sinorama (Group, 2013): a news magazine with 9.87 million words; (ii) CIRB030 (LIPS & Labs, 2013): a test bench for information retrieval consisting of several domains with 124.4 million words; and (iii) Sinica Corpus (ACLCLP, 2013b): a general text corpus collected for language analysis with 4.8 million words. A conditional random field (CRF)-based tagger (Huang, Chiang, Wang, Yu, & Chen, 2010; Lafferty, McCallum, & Pereira, 2001) was employed to segment the corpus into word/part of speech (POS) sequences.

A 60k-word lexicon was also constructed based on the word frequency.

The prosodic models were trained using data from 164 speakers (about 8.3 hours) extracted from the TCC300 training set by a sequential optimization algorithm. A subset of the TCC300 training set was adopted as development set in order to determine the weighting vector for model combination. This development set covered utterances by 33 speakers and with a minimum length of 18 minutes. The parameters of the attribute detectors were estimated using part of the SI-84 set of the Wall Street Journal Corpus (WSJ0; Paul & Baker, 1992) with 7,077 utterances by 84 speakers, or 15.3 hours of speech. A cross-validation (CV) set was generated by extracting 200 sentences out of the SI84 training set. The CV set accounts for about 3% of the SI-84 set and was used to terminate the training. The remaining 6,877 SI-84 sentences were used as training material. The word lattices for rescoring were generated by HTK 3.4.1 with the tri-gram LM and MMI-trained syllable models. The word coverage rate of the lattice was 93.75%, which is in the top band of performance for the proposed approach to attain.

Experimental results are listed in Table 1, from which it can be seen that the baseline performance can be improved by incorporating knowledge sources. The best performance for word, character and syllable recognition can be achieved by combining scores of manner (+M), break type (+B), and prosodic state information (+P). Generally, the break-type information alone could provide a better improvement on WER/CER/SER than the manner attribute.

Table 1: Comparison of word, character and syllable error rates (in %; WER, CER and SER, respectively) in Mandarin speech recognition by adding various knowledge sources, including manner (+M), break type (+B) and prosodic states (+P) to the baseline ASR system.

	WER	CER	SER
Baseline	13.75	10.56	7.79
+M	13.45	10.20	7.44
+B	12.57	9.81	7.41
+M+B	12.43	9.36	6.90
+B+P	12.26	8.93	4.73
+M+B+P	12.24	8.55	6.63

The relatively limited enhancement obtained with the manner information was mainly due to the fact that only cross-language attribute detectors were used. The lack of exact phone boundaries within a syllable also negatively affects the attribute accuracy.

4. CONCLUSION: CHALLENGES TO THE SPEECH SCIENCE COMMUNITY

The design of state-of-the-art top-down ASR systems is based on large amounts of language-specific, quality-controlled data and pronunciation dictionaries that represent words as sequences of pre-defined sound units. As previously mentioned, this conventional approach fails whenever a new application must be developed when speakers with a non-native accent or with some speech impairment must be recognized, or when speech that was produced in a noisy environment must be recognized.

It can be argued that humans do not need enormous amounts of training data to be able to understand new speakers or cope with new acoustic backgrounds. Thus, the conventional top-down ASR approach that relies on training generative models on the basis of enormously large amounts of data misses essential aspects of the structure of speech signals that allow humans to substantially outperform automatic systems with only a fraction of the training data. Furthermore, attempts to train language-independent ASR models using such top-down approaches have failed to adequately capture variability in speech. In this paper, we have argued that the missing piece of the puzzle lies in designing a bank of "perfect" feature detectors. These speech detectors should be stochastic in nature, and the data-driven modeling techniques in state-of-the-art systems can be extended to a bottom-up detection approach to ASR in which speech feature detection and linguistic knowledge integration play key roles. We have referred to this bottom-up approach as ASAT.

ASAT is by design a collaborative research framework that proposes a radical departure from traditional top-down ASR and proposes a bottom-up data-driven approach with the goal of closing the gap between HSR and ASR. An important aspect of the detection-based ASAT paradigm is that it gives the opportunity to better understand the flaws in the ASR recognizer. For example, if the /p/ sound is systematically confused with the /b/ sound, this may show that the voicing detector needs to be improved. Moreover, the detection-based approaches inherently provide a platform in which expert knowledge of linguistic and acoustic phonetics can be methodically incorporated into the system. The collection of information includes a set of fundamental speech sounds and their linguistic interpretations, a speaker profile that encompasses gender, accent and other speaker characteristics, the speaking environment that describes the interaction between speech and acoustics, and many other speech characterizations.

ASAT would allow different groups working on different components of the system to improve the overall system performance, because there are many pieces of information, or acoustic cues, to be extracted and utilized. In the meantime modular approaches are usually more computationally tractable than integrated approaches. Furthermore, 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 landmarks, such as vowels, glides and fricatives, in adverse conditions. They could prove useful for selecting good data segments and designing signal-specific speech en-

hancement, feature compensation, and model adaptation algorithms for reliable information extraction.

In summary, it may be 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 for other modules related to this attribute, and probably for the overall system. 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 collectively incorporated into the "best" overall next generation speech processing system.

REFERENCES

Association for Computational Linguistics and Chinese Language Processing (ACLCLP) (2013a). *Mandarin microphone speech corpus–TCC300* [Database]. Retrieved from http://www.aclclp.org.tw/use_mat.php#tcc300edu

Association for Computational Linguistics and Chinese Language Processing (ACLCLP) (2013b). *Sinica balanced corpus* (version 4.0) [Corpus]. Retrieved from http://www.aclclp.org.tw/use_asbc.php

Bahl, L. R., Brown, P. F., de Souza, P. V., & Mercer, R. L. (1986). Maximum mutual information estimation of HMM parameters for speech recognition. *Proceedings of the 1986 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '86), 11*, 49–512. http://dx.doi.org/10.1109/ICASSP.1986.1169179

Baker, J. (1975). The DRAGON system–An overview. *IEEE Transactions on Acoustics, Speech and Signal Processing, 23*(1), 24–29. http://dx.doi.org/10.1109/TASSP.1975.1162650

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities, 3*, 1–8.

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics, 37*(6), 1554–1563. http://dx.doi.org/10.1214/aoms/1177699147

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics, 41*(1), 164–171. http://dx.doi.org/10.1214/aoms/1177697196

Bellegarda, J. R. (2000). Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE, 88*(8), 1279–1296. http://dx.doi.org/10.1109/5.880084

Bellman, R. E. (1957). *Dynamic programming.* Princeton, NJ: Princeton University Press.

Bellman, R. E., & Dreyfus, S. E. (1962). *Applied dynamic programming*. Princeton, NJ: Princeton University Press.

Chiang, C.-Y., Siniscalchi, S. M., Wang, Y.-R., Chen, S.-H., & Lee, C.-H. (2012). A study on cross-language knowledge integration in Mandarin LVCSR. *Proceedings of the 2012 8th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 315–319. http://dx.doi.org/10.1109/ISCSLP.2012.6423528

De Mori, R. (Ed.). (1998). *Spoken dialogues with computers*. San Diego, CA: Academic Press.

Deng, L., & Yu, D. (2011). Deep convex network: A scalable architecture for speech pattern classification. *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH '11)*, 2285–2288.

Fousek, P., & Hermansky, H. (2006). Towards ASR based on hierarchical posterior-based keyword recognition. *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06), 1*, 433–436. http://dx.doi.org/10.1109/ICASSP.2006.1660050

Gauvain, J.-L., & Lamel, L. (2000). Large vocabulary continuous speech recognition: Advances and applications. *Proceedings of the IEEE, 88*(8), 1181–1200. http://dx.doi.org/10.1109/5.880079

Group, C. W. (2013). *Taiwan Panorama Magazine text corpus*. Retrieved from http://www.aclclp.org.tw/use_ gh_c.php

Hacioglu, K., Pellom, B., & Ward, W. (2004). Parsing speech into articulatory events. *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04), 1*, 925–928. http://dx.doi.org/10.1109/ICASSP.2004.1326138

Hasegawa-Johnson, M., Baker, J., Borys, S., Chen, K., Coogan, E., Greenberg, ... Wang, T. (2005). Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop. *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05),* 213–216. http://dx.doi.org/10.1109/ICASSP.2005.1415088

Hou, J. (2009). *On the use of frame and segment-based methods for the detection and classification of speech sounds and features*. Doctoral dissertation. Rutgers University, NJ, USA.

Hou, J., Rabiner, L. R., & Dusan, S. (2007). On the use of time-delay neural networks for highly accurate classification of stop consonants. *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH '07)*, 1929–1932.

Huang, Q.-Q., Chiang, C.Y., Wang, Y.-R., Yu, H.-M., & Chen, S.H. (2010). Variable speech rate Mandarin Chinese text-to-speech system. *Proceedings of the 22th Conference on Computational Linguistics and Speech Processing (ROCLING '10*), 222–235.

Jelinek, F. (1997). *Statistical method for speech recognition*. Cambridge, MA: The MIT Press.

Jeon, W., & Juang, B. H. (2007). Speech analysis in a model of the central auditory system. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(6), 1802–1817. http://dx.doi.org/10.1109/TASL.2007.900102

Kawahara, T., Lee, C.H., & Juang, B.-H. (1998). Flexible speech understanding based on combined key-phrase detection and verification. *IEEE Transactions on Speech and Audio Processing, 6*(6), 558–568. http://dx.doi.org/10.1109/89.725322

Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01),* 282–289.

Lee, C.-H. (2003). On automatic speech recognition at the dawn of the 21st century. *IEICE Transactions on Information and Systems*, *86*(3), 377–396.

Lee, C.-H., Clements, M. A., Dusan, S., Fosler-Lussier, E., Johnson, K., Juang, B.-H., & Rabiner, L. R. (2007). An overview on automatic speech attribute transcription (ASAT). *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH '07),* 1825–1828.

Lee, C.-H., & Huo, Q. (2000). On adaptive decision rules and decision parameter adaptation for automatic speech recognition. *Proceedings of the IEEE, 88*(8), 1241–1269. http://dx.doi.org/10.1109/5.880082

Lee, C.-H., & Siniscalchi, S. M. (2013). An information-extraction approach to speech processing: Analysis, detection, verification, and recognition. *Proceedings of the IEEE, 101*(5), 1089–1115. http://dx.doi.org/10.1109/JPROC.2013.2238591

Lee, C.-H., Soong, F. K., & Paliwal, K. K. (Eds.). (1996). *Automatic speech and speaker recognition: Advanced topics*. Boston, MA: Kluwer Academic. http://dx.doi.org/10.1007/978-1-4613-1367-0

Lehmann, E. L. (1959). *Testing statistical hypotheses*. New York, NY: Wiley. http://dx.doi.org/10.1007/978-1-4757-1923-9

Li, J., & Lee, C.-H. (2005). On designing and evaluating speech event detectors. *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05–EUROSPEECH '05)*, 3365–3368.

Li, J., Tsao, J., & Lee, C.-H. (2005). A study on knowledge source integration for candidate rescoring in automatic speech recognition. *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05), 1*, 837–840. http://www.dx.doi.org/10.1109/ICASSP.2005.1415244

LIPS & Labs, N. (2013). Chinese Information Retrieval Benchmark (CIRB030) (Version 3.0) [test collection]. Retrieved from http://www. aclclp.org.tw/use_cir.php

Liu, S. A. (1996). Landmark detection for distinctive feature-based speech recognition. *Journal of the Acoustical Society of America, 100*(5), 3417–3430. http://dx.doi.org/10.1121/1.416983

Ma, C., & Lee, C.-H. (2007). A study on word detector design and knowledge-based pruning and rescoring. *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH '07)*, 1473–1476.

Mohamed, A. R., Dahl, G., & Hinton, G. E. (2009, December). *Deep Belief Networks for phone recognition*. NIPS Workshop on Deep Learning for Speech Recognition and Related Applications, Whistler, BC, Canada.

Mohri, M. (1997). Finite-state transducers in language and speech processing. *Computational Linguistics, 23*(2), 269–311.

Morris, J., & Folser-Lussier, E. (2006). Combining phonetic attributes using conditional random fields. *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH '06)*, 597–600.

Ney, H., & Ortmanns, S. (2000). Progresses in dynamic programming search for LVCSR. *Proceedings of the IEEE, 88*(8), 1224–1240. http://dx.doi.org/10.1109/5.880081

Niyogi, P., Mitra, P., & Sondhi, M. (1998). A detection framework for locating phonetic events. *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP '98)*, paper 0665.

O'Shaughnessy, D. (2000). *Speech communication: Human and machine*. IEEE Press.

Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the Workshop on Speech and Natural Language* (pp. 357–362). Stroudsburg, PA: Association for Computational Linguistics. http://dx.doi.org/10.3115/1075527.1075614

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected application in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286. http://dx.doi.org/10.1109/5.18626

Rabiner, L. R., & Juang, B.-H. (1993). *Fundamentals of speech recognition*. Prentice Hall.

Rabiner, L. R., & Schafer, R. W. (2011). *Theory and applications of digital speech processing*. Pearson Higher Education.

Ramesh, P. and Niyogi, P. (1998). The voice feature for stop consonants: Acoustic phonetic analysis and automatic speech recognition experiments. *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP '98)*, paper 0881.

Seide, F., Li, G., & Yu, D. (2011). Conversational speech transcription using context-dependent deep neural networks. *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH '11)*, 437–440.

Shamma, S. (2001). On the role of space and time in auditory processing. *Trends in Cognitive Sciences, 5*(8), 340–348.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379–423. http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x. Ibid., *27*(4), 623–656. http://dx.doi.org/10.1002/j.1538-7305.1948.tb00917.x

Siniscalchi, S. M., & Lee, C.-H. (2009). A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition. *Speech Communication, 51*(11), 1139–1153. http://dx.doi.org/10.1016/j.specom.2009.05.004

Siniscalchi, S. M., Li, J., & Lee, C.-H. (2006). A study on lattice rescoring with knowledge scores for automatic speech recognition. *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH '06)*, 517–520.

Siniscalchi, S. M., Lyu, D.-C., Svendsen, T., & Lee, C.-H. (2012). Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(3), 875–887. http://dx.doi.org/10.1109/TASL.2011.2167610

Siniscalchi, S. M., Reed, J., Svendsen, T., & Lee, C.-H. (2013). Universal attribute characterization of spoken languages for automatic spoken language recognition. *Computer Speech & Language, 27*(1), 209–227. http://dx.doi.org/10.1016/j.csl.2012.05.001

Siniscalchi, S. M., Svendsen, T., & Lee, C.-H. (2009). A phonetic feature based lattice rescoring approach to LVCSR. *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '09),* 3865–3868. http://dx.doi.org/10.1109/ICASSP.2009.4960471

Siniscalchi, S. M., Svendsen, T., & Lee, C.-H. (2011). A bottom-up stepwise knowledge-integration approach to large vocabulary continuous speech recognition using weighted finite state machines. *Proceedings of the 12th Annual Conference of the International SpeechCommunicationAssociation(INTERSPEECH'11)*,901–904.

Stolcke, A. (2002). SRILM–An extensible language modeling toolkit. *Proceedings of the 7th Conference on Spoken Language Processing (ICSLP '02–INTERSPEECH '02)*, 16–20.

Sukkar, R. A., & Lee, C.-H. (1996). Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition. *IEEE Transactions on Speech and Audio Processing, 4*(6), 420–429. http://dx.doi.org/10.1109/89.544527

Tsao, Y., Li, J., & Lee, C. H. (2005). A study on separation between acoustic models and its applications. *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05–EUROSPEECH '05)*, 1109–1112.

Yu, D., Siniscalchi, S. M., Deng, L., & Lee, C.-H. (2012). Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition. *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '12),* 4169–4172. http://dx.doi.org/10.1109/ICASSP.2012.6288837

Zue, V. W. (1981). Acoustic-phonetic knowledge representation: Implications from spectrograms reading experiments. In J.-P. Jaton (Ed.), *Automatic speech analysis and recognition* (pp. 101–120). http://dx.doi.org/10.1007/978-94-009-7879-9_5