A Bottom-Up Modular Search Approach to Large Vocabulary Continuous Speech Recognition

Sabato Marco Siniscalchi, *Member, IEEE*, Torbjørn Svendsen, *Senior Member, IEEE*, and Chin-Hui Lee, *Fellow, IEEE*

***Abstract*—A novel *bottom-up* decoding framework for large vocabulary continuous speech recognition (LVCSR) with a modular search strategy is presented. Weighted finite state machines (WFSMs) are utilized to accomplish stage-by-stage acoustic-to-linguistic mappings from low-level speech attributes to high-level linguistic units in a bottom-up manner. Probabilistic attribute and phone lattices are used as intermediate vehicles to facilitate knowledge integration at different levels of the speech knowledge hierarchy. The final decoded sentence is obtained by performing lexical access and applying syntactical constraints. Two key factors are critical to warrant a high recognition accuracy, namely: (i) generation of high-precision sets of competing hypotheses at every intermediate stage; and (ii) low-error pruning of unlikely theories to reduce input lattice sizes while maintaining high-quality hypotheses for the next layers of knowledge integration. The decoupled nature of the proposed techniques allows us to obtain recognition results at all stages, including attribute, phone and word levels, and enables an integration of various knowledge sources not easily done in the state-of-the-art hidden Markov model (HMM) systems based on top-down knowledge integration. Evaluation on the Nov92 test set of the 5000-word, Wall Street Journal task demonstrates that high-accuracy attribute and phone classification can be attained. As for word recognition, the proposed WFSM-based framework achieves encouraging word error rates. Finally, by combining attribute scores with the conventional HMM likelihood scores and re-ordering the N-best lists obtained from the word lattices generated with the proposed WFSM system, the word error rate (WER) can be further reduced.**

***Index Terms*—Artificial neural network, knowledge integration, large vocabulary continuous speech recognition (LVCSR), speech attribute detection, weighted finite state machines (WFSM).**

I. Introduction

STATE-OF-THE-ART automatic speech recognition (ASR) technology is based on a pattern matching framework that is motivated by expressing spoken utterances as stochastic patterns [1]. Hidden Markov models (HMMs) (e.g., [2]) have then been used to characterize these speech patterns, from phones to syllables, words and sentences. A single finite state network (FSN), composed of acoustic HMM states of grammar nodes and their connecting arcs [3], is then constructed to represent all ASR task constraints, known as *top-down* knowledge integration. For a given input utterance ASR is performed by searching the FSN via *dynamic programming* (DP) based optimal decoding (e.g., [4]) to obtain the most likely sequence of words as the recognized sentence using maximum a posteriori (MAP) decoding (e.g., [5], [6]). We will refer to this type of decoding strategy as *integrated search*.

This statistical pattern matching approach to ASR relies on collecting a large amount of speech and text examples and learning the HMM parameters without the need to use detailed knowledge about a target language. It offers an advantage for automatic model learning from data via a rigorous mathematical formulation. We have witnessed almost four decades on three major HMM technology advances, namely: (i) *detailed modeling* – capable of characterizing thousands of context-dependent phone units with millions of parameters using publicly available software packages (e.g., HTK [7]); (ii) *adaptive modeling* – capable of learning an unseen acoustic condition with a small amount of condition-specific adaptation data (e.g., [6], [8]–[10]); and (iii) *discriminative* modeling – capable of obtaining HMMs that are discriminative among competing unit models (e.g., [11]–[15]).

On the other hand, speech researchers would agree that the ASR problem is still far from solved due to the degrading performance of the state-of-the-art ASR systems in mismatch training and testing conditions. Furthermore, poor accuracies are observed when dealing with spontaneous speech, where ill-formed utterances are usually encountered. It is worth noting that the word error rate (WER) on the Switchboard task [16] has been reduced to below 20% only very recently [17], and yet this level of performance is still rather poor when compared with LVCSR tasks of a similar complexity, e.g., the Wall Street Journal (WSJ) task [18].

In order to mitigate some of the ASR limitations, we have seen the utilization of knowledge sources in speech production (e.g., [19], [20]) and auditory processing and perception (e.g., [21]–[23]. Many of them are not easily integrated into the conventional top-down ASR systems. The need for alternative ASR paradigms that are capable of leveraging on existing speech literature has thus attracted some research attention in recent years, and a few significant examples closely related to our work will be briefly reviewed in Section III. Most of these attempts

Manuscript received July 02, 2012; revised September 29, 2012; accepted December 08, 2012. Date of publication December 20, 2012; date of current version January 18, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Haizhou Li.

S. M. Siniscalchi is with the Faculty of Architecture and Engineering, University of Enna "Kore," 94100 Enna, Italy (e.mail: marco.siniscalchi@unikore.it).

T. Svendsen is with the Department of Electronics and Telecommunications, Norwegian University of Science and Technology, 7491 Trondheim, Norway (e-mail: torbjorn@iet.ntnu.no).

C.-H. Lee is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: chl@ece.gatech.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2012.2234115

focused on detecting speech cues or acoustic landmarks that are usually utilized in a bottom-up manner. The process 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 into the overall system design.

In this paper, we present a decoding framework that employs a *decoupled search* strategy as opposed to the delayed decision, integrated search approach. ASR is accomplished in a bottom-up, stage-by-stage fashion by performing back-end lexical access and syntax knowledge integration over the output of a high-accuracy phone-based front-end, which generates frame-level speech attribute detection scores and phone posterior probabilities. The AT&T WFSM [24] package has been used to implement the proposed decoding framework, because it offers a flexible architecture that will ease future integration of more complex types of knowledge sources. Other similar tools, such as OpenFST [25], can also be utilized.

Two key factors are critical to warrant a high recognition accuracy using a bottom-up modular search strategy, namely: (i) high-precision phone lattice generation at every intermediate high-accuracy front-end; and (ii) low-error pruning of unlikely theories to reduce input lattice sizes while maintaining high-quality hypotheses for the next stages of knowledge integration. In the current implementation we use multi-layer perceptrons (MLPs) to realize both attribute detection, phone classification, and attribute-to-phone mapping for phone lattice generation. We also use probability scores at the attribute, phone, and word levels to perform lattice pruning.

We evaluated the proposed technique on the 5000-word WSJ task using the Nov92 test set of 330 test utterances. Earlier findings were reported in [26]. We have observed that the proposed approach outperforms the *TANDEM* ASR system [27] where phone probability vectors generated by an ANN are used as normal feature vectors in a conventional HMM-based system. Furthermore, our approach attains superior performance to various connectionist LVCSR [28] evaluated on the same task [29]. Finally, by combining attribute scores with the conventional HMM likelihood scores and re-ordering the N-best lists obtained from the word lattices generated with the proposed WFSM system, the resulting WERs are lower than the conventional HMM-based systems trained with either the ML or MMI criterion without taking advantage of the additional information provided by the high-precision speech attribute detectors and the high-quality word lattices.

The remainder of the paper is organized as follows: Section II compares the traditional recognition framework with the bottom-up modular paradigm. Section III provides a brief survey of recent alternative ASR paradigms related to our work. Section IV describes the proposed bottom-up, decoupled LVCSR system. The two key modules, the high-accuracy front-end and the knowledge-integration back-end, are discussed in detail. Next, we present the experimental setup and report the experimental results in Section V. A discussion on some limitations of the currently prevailing integrated search and potential remedies offered by the proposed bottom-up modular search strategy is given in Section VI. Finally we summarize our findings in Section VII.

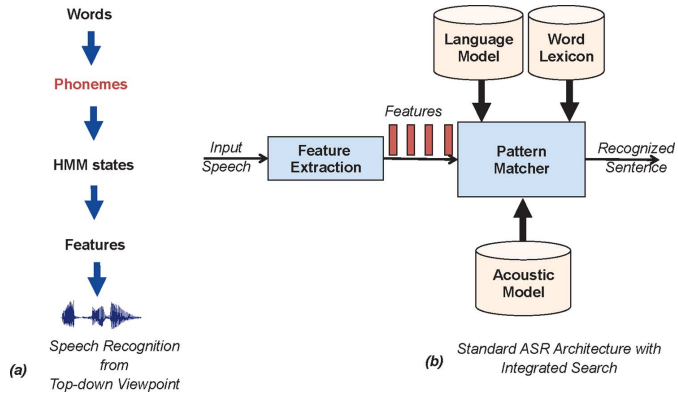

Fig. 1. A block diagram of a typical top-down approach to LVCSR with an integrated search strategy.

II. Top-Down vs. Bottom-Up LVCSR

Fig. 1(a) gives a pictorial representation of how spoken utterances are formed under a top-down ASR paradigm. Hence a sequence of words is seen as being composed of a sequence of phoneme units, which are in turn composed of a sequence of HMM states. HMMs have the nice property that they can simultaneously characterize spectral and temporal variations of the acoustic feature vectors. The notion that a word is composed of a sequence of phonemes is often referred to as "beads-on-a-string" [30], [31].

A block diagram of the top-down ASR system with an integrated search is shown in Fig. 1(b). The feature extraction module typically performs time-synchronous cepstral analysis. The "Pattern Matcher" block generates a finite state network representation of all key task constraints, such as lexicon and language model. In principle, this search strategy achieves the highest performance if all the knowledge sources can be completely characterized and fully integrated using the speech knowledge hierarchy in the linguistic structure of acoustics, lexicon, syntax and semantics.

On the other hand, the compiled FSN for LVCSR tasks is often too large, and it therefore becomes computationally expensive to find the best sentence through a huge and ever-expanding search space. Thus all knowledge sources must be kept simple in order to be efficiently combined into a single search space. This has particularly inhibited progress at the linguistic level, and almost all LVCSR systems employ non-optimal linguistic components such as static lexicons (lexicalization of morphological processes) and n-gram language models (LMs) which force the decoding process to generate hypotheses which sometimes conflict with the acoustic constraints as shown in [32].

The top-down integrated framework therefore often hampers the definition of generic knowledge sources that can be used in different domains. As a result, applications for a new knowledge domain need to be built almost from scratch. In addition, the effectiveness of the integrated search diminishes when dealing with unconstrained speech input, since more complex language models are needed for handling spontaneous speech, which often includes extraneous words, hesitations, repetitions and unexpected expressions. In some cases, we even can observe many utterances that are totally irrelevant to the

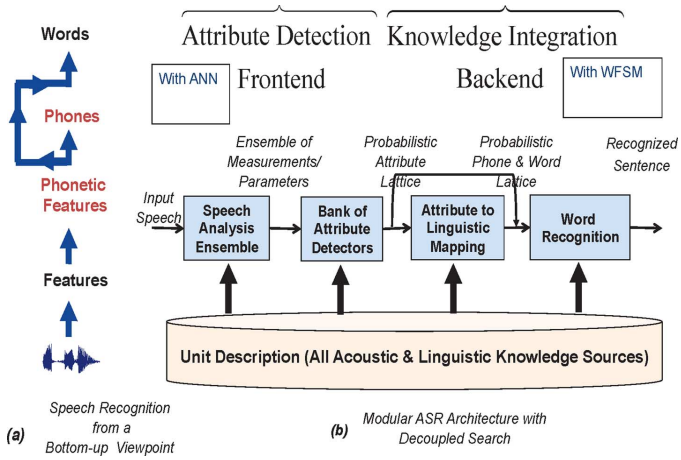

Fig. 2. A block diagram of the proposed bottom-up LVCSR with a decoupled modular search strategy.

task. These kinds of spoken utterances cannot be completely characterized even if a large amount of training speech data is available.

For the bottom-up decoupled approach to ASR shown in Fig. 2(a), the recognized sentence can be obtained through sequential refinements starting from the raw speech signal up to the sequence of words. As long as the interface between the adjacent decoding modules can be completely specified, each module can be designed and tested separately. It should be pointed out that a word could in theory be directly expressed in terms of phonetic features, which are more fundamental units than phonemes [33]. This solution would allow us to overcome the main issue of the "beads-on-a-string" paradigm that forces a sequence of acoustic observations to be synchronized with a sequence of phonemes.

A block diagram of the proposed bottom-up modular LVCSR system is illustrated in Fig. 2(b) in which the front-end is realized with a speech analysis ensemble to produce all the speech parameters and a bank of attribute detectors intended to spot all the relevant speech cues and events to be passed on for the next stage of knowledge integration. ASR is accomplished in a bottom-up fashion by performing back-end lexical access and syntax knowledge integration over the output of our high-accuracy front-end, which generates frame-level speech attribute detection scores and phone posterior probabilities.

Multi-stage recognition has been proposed in the past, e.g., [34], [35]. Nonetheless simple versions of all knowledge sources are using the initial decoding step, and an integrated search is adopted. A rescoring mechanism is then employed to obtain better recognition results by using more complex models. In contrast, the key idea of the proposed modular search strategy is to use the best models available at each recognition stage (e.g., best nasal detector) in order to minimize information loss. We believe that modularity can allow more accurate and less domain dependent recognition at the expense of recognition speed.

III. Related Alternative Approaches to ASR

In this section we provide a brief review of some alternative ASR approaches that are related to bottom-up LVCSR. In [36], good performance was demonstrated in a speech understanding task by using key phrase detection followed by utterance verification. Specifically, this framework was able to maintain a good accuracy for well-formed utterances while performing much better than integrated search approaches for ill-formed utterances. In [37], a template-based approach to ASR was proposed. However a good template selection procedure is required in order to keep the computational load within bounds. New theories of nonlinear phonology, articulatory phonology, and landmark-based speech perception were employed in [38] to design a segment-based, multi-stage recognizer.

In [39], three landmark-based ASR systems were proposed and differed by the pronunciation model used. Several goals, representing a step forward in knowledge-based approach to ASR, had been achieved in that study. For example, the proposed two-stage algorithm can find prominent landmark differences among words in different hypotheses and trigger a selected number of landmark detectors that help to choose the best theory. Unfortunately, the LVCSR results were not as compelling as expected.

The FlaVoR decoder proposed in [40] employed a modular architecture to tackle ASR. The system is a two-layered architecture where the first layer is a phone decoder that generates a dense phone network, and the second layer is a robust decoder that finds those words from the lexicon that match well with the phone sequences encoded in the phone network. The acoustic models used in FlaVoR were context-dependent tied-state phoneme HMMs. Furthermore, morpho-phonological and a morpho-syntactic knowledge were fused in a single finite-state transducer.

It is interesting to notice that FlaVoR and the proposed bottom-up modular approach share the same goal: to abandon the standard all-in-one search strategy with integrated acoustic, lexical and language model information in favor of a modular search framework which allows for the integration of more complex linguistic components, especially with unconstrained LVCSR [41]. FlaVoR tried to combine the advantages of top-down and bottom-up strategies using a single framework where the two paradigms meet at some convenient or intermediate levels in which a phone graph can be generated. Nonetheless, a finer stage-by-stage search technique is implemented in our current study where more complex models and additional information are used to refine the search space in subsequent passes.

In [42], the authors argued that speech features and lexical words are inherently correlated in natural language and demonstrated better recognition accuracies can be obtained by jointly optimizing acoustic and linguistic parameters according to the maximum entropy principle. Conditional random fields (CRFs) [43] are a mathematical tool that can also be pursue joint optimization of independent features, and a recent CRF approach to LVCSR [44], [45] with a companion toolkit called SCARF [46] was developed for performing ASR with segmental CRFs.

The recently proposed automatic speech attribute transcription (ASAT) framework [47], [48] views ASR from bottom-up in a "divide-and-conquer" perspective and aims at identifying acoustic and linguistic information not fully exploited by the current top-down ASR paradigm. Our bottom-up approach to

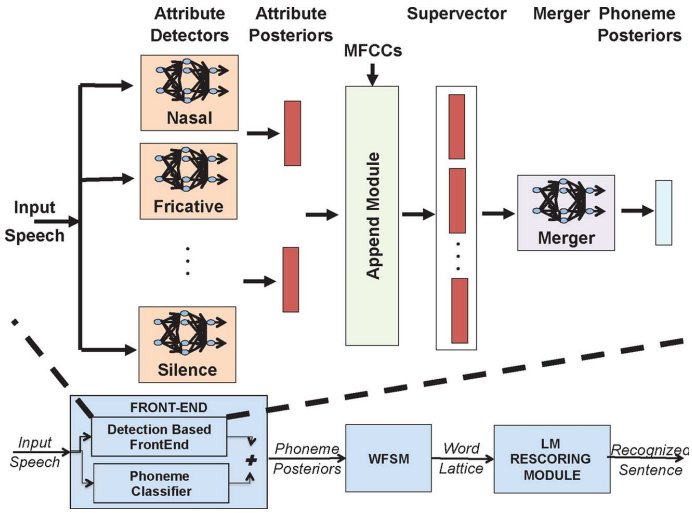

Fig. 3. The overall bottom-up LVCSR system with decoupled search strategy. The front-end generates phone posterior probabilities which are fed into the WFSM module at a word level, and then rescored using a trigram LM.

LVCSR, which was derived from ASAT, will be presented in detail in the following.

IV. Overall Bottom-Up LVCSR System

The proposed WFSM-based system consists of three main modules: (i) a high-accuracy front-end that scores speech frames and generates phoneme posterior probabilities, (ii) a module that combines these scores and generates a word-level lattice, and (iii) a language model rescoring block. The first module is practically a phone classifier based on two components (i) an ensemble of speech attribute detectors, each of which independently analyzes input speech and provides a set of detection scores, and a merging block that combines these detection scores and generates phone posterior probabilities, and (ii) a hierarchical structure of MLPs that generates phoneme posterior probabilities for each input speech. The second module is a chain of WFSMs that generates a word-level lattice, later referred as *WFSM-lattice*, by imposing duration, lexical, and grammar constraints. In the future, possible additional information available at each step can be integrated with ease. Complex models can also be employed. The final block is an LM rescoring module that integrates trigram LM information and produces the N most likely sentence hypotheses, or N-best list. The overall system is a bottom-up, stepwise word decoder, as shown in Fig. 3.

A. Front-End

The front-end shown in the lower part of Fig. 3 consists of two blocks: a detection-based front-end that scores speech attributes and generates phoneme posterior probabilities, and a phoneme classifier based on a hierarchical structure of three MLPs.

The upper part of Fig. 3 shows the detection-based front-end used in this study. It consists of two main blocks: (i) a bank of speech attribute detectors that provides attribute posterior probabilities; and (ii) an event merger that maps attribute posteriors into phoneme posteriors. An attribute detector was built for each of the following 20 phonetic features: fricative, approximant, nasal, stop, vowel, coronal, dental, glottal, high, labial, low, mid, retroflex, velar, anterior, back, continuant, round, tense, and voiced. Furthermore, a detector for silence was also designed. Each detector analyzes a frame of the input speech signal and classifies a given input into "attribute present" or "attribute absent" outputs. The bank of attribute detectors was implemented using a hierarchical structure of MLPs. The goal is to exploit long temporal dependencies in the speech signal, since that has been proven beneficial in several classification tasks, e.g., [49]–[52]. This will also help the attribute and posterior models because they are all *context-independent* in contrast to state-of-the-art acoustic triphone models.

Long temporal information is taken into account by extracting sub-band energy trajectories with a 23-band uniform mel-frequency filterbank. For each critical band a window of 310 *ms* centered around the frame being processed is considered and split in two halves: left-context and right-context (referred to as split temporal context features) [50]. Two independent front-end MLPs ("lower nets") are trained on those two halves and generate left- and right-context speech attribute posterior probabilities, respectively. Usually, the discrete cosine transform is applied to the input of these lower nets to reduce the dimensionality (see [50] for more details). The outputs of the two lower nets are then sent to the third MLP that acts as a merger and gives the attribute posterior probability of the target speech attribute. Feed-forward MLPs were used with a single hidden layer having 500 hidden units.

The "Append Module" stacks together the outputs delivered by the attribute detectors along with a 39-dimensional vector of Mel-frequency cepstrum coefficients for a given input frame and generates a supervector. The "Merger" is then trained using these supervectors to generate posterior probabilities at a phone level. A frame-based MLP with a single hidden layer having 1500 hidden units was used to implement the "Merger." This MLP discriminates between 40 phoneme classes.

The "Phoneme Classifier" shown in the lower part of Fig. 3 is implemented with a hierarchical structure of three MLPs trained over long temporal features generated as described above. Specifically, two independent front-end MLPs ("lower nets") is trained on the left- and right-context of the split temporal context features to generate left- and right-context speech phoneme posterior probabilities, respectively. The outputs of the two lower nets are then sent to a third MLP that acts as a merger and gives the phone posterior probability of the target speech attribute. Each MLP has a single-hidden layer having 800 hidden units. The "Phoneme Classifier" is directly trained on acoustic features rather than on attribute scores like the "Merger."

The phoneme posterior probabilities at the output of the "FRONT-END" module are compted using a linear combination rule. Specifically, given two classifiers and V classes $c_1, \ldots c_V$, the sum rule used to combine these two classifiers,

$$P(c_k|x) = \alpha P^{(1)}(c_k|x) + \beta P^{(2)}(c_k|x), \quad k = 1, \ldots, V$$
$$\alpha + \beta = 1 \quad \alpha \geq 0, \quad \beta \geq 0. \tag{1}$$

Here α and β denote the interpolation coefficients, and $P^i(c_k|x)$ is the phone posterior probability for class c_k esti-

mated by the ith classifiers. In the following experiments, α and β were set to 0.5.

In this paper, the sigmoid and softmax non-linearities are used throughout as activation functions in the hidden and output nodes, respectively. The standard back propagation algorithm [53] is adopted as the training method, and to avoid over-fitting, the reduction in classification error on a development set during the training phase is chosen as a stopping criterion.

The ICSI QuickNet neural network software package[1] is used to implement both attribute detectors and phoneme classifiers.

B. Knowledge Integration Back-End

LVCSR is accomplished by bottom-up knowledge integration, starting with the frame-level evidence gathered at the output of the front-end at the bottom level represented as an acceptor ($\mathcal{F}$). In practice, $\mathcal{F}$ is a graph with a number of states that equals the length of the input sentence (in frames), and a number of edges between each consecutive pair of states that equals the output dimension of the merger (i.e., the number of events/phonemes to be classified). Each transition edge carries a weight that is equal to $-\log_e(Prob(Phoneme\,|\,Speech\,Input)/Prior(Phoneme))$. In this work, we assumed all phonemes have the same prior probability. Thus, each arc carries a cost represented as a positive number. A duration transducer $\mathcal{D}$ is a three-state FSM that maps sequences of frames into a single phoneme symbol. $\mathcal{D}$ is composed with the feature acceptor $\mathcal{F}$ in order to ensure a minimum duration of three frames for each single phone unit, and a recognition network $\mathcal{FD}$ is obtained after composition. The weights of $\mathcal{D}$ are generated by the conventional ML estimation procedure usually adopted to train the transition matrix of a conventional HMM-based LVCSR system [54]. Word recognition can now be performed by lexical access to a vocabulary of words along with integration of syntax knowledge. These two tasks are accomplished by composing the $\mathcal{FD}$ transducer with a lexicon transducer ($\mathcal{L}$) and a back-off language acceptor ($\mathcal{G}$).

A recognition network ($\mathcal{RN}$) that maps frame distributions on the input side to word strings on the output side can be obtained as follows:

$$\mathcal{RN} = \mathrm{Min}(Det(\mathcal{F}\,o\,\mathcal{D}\,o\,\mathcal{L}\,o\,\mathcal{G})) \tag{2}$$

where o indicates the composition operation. Det indicates the determinization procedure, and Min refers to the minimization process. The WFSM paradigm can thus be utilized to combine all levels of knowledge sources in our proposed LVCSR system into an integrated $\mathcal{RN}$.

As mentioned earlier, the advantages of introducing modularity in the ASR engine come with a cost, namely a slow recognition speed and a large memory requirement. To address these issues, a decoding strategy with a decoupled knowledge integration scheme is implemented as follows:

$$d\mathcal{RN} = (((\mathcal{F}'\,o\,\mathcal{D})'\,o\,\mathcal{L})'\,o\,\mathcal{G}_{bigram})' \tag{3}$$

where the prime symbol ($'$) indicates that a pruning operation has been performed before composition, and the parentheses mark the composition stages. Minimization and determinization were performed although not shown in (3). The decoupled $\mathcal{RN}$ ($d\mathcal{RN}$) was then rescored with a trigram LM. The pruning process at the output of each composition step is performed to remove highly unlikely search paths and thus reduce the size of the intermediate recognition networks. Because of this operation some correct states and paths may be removed before the next stage of knowledge composition and cause search and recognition errors. Care is needed to minimize such problems.

[1]ICSI quicknet software package, http://www.icsi.berkeley.edu/speech/qn.htm

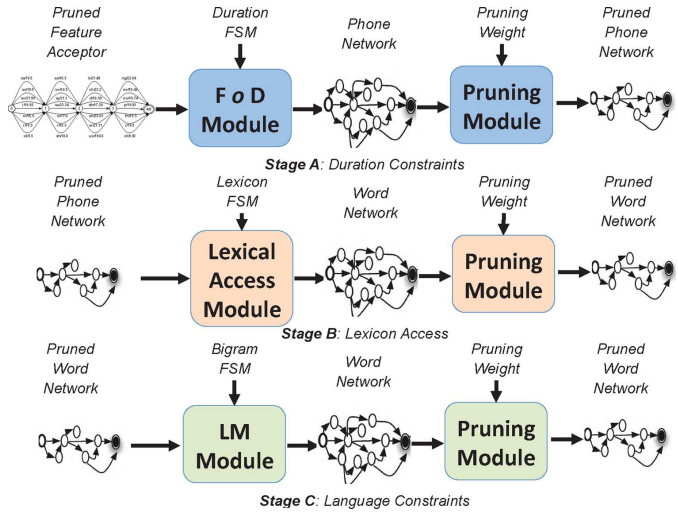

Fig. 4. Decoupled, bottom-up, detection-based LVCSR with a bigram LM.

C. LVCSR Implementation

Fig. 4 shows how the decoupled approach is implemented in practice using WFSMs. The original feature acceptor $\mathcal{F}$ is pruned and composed with the duration transducer $\mathcal{D}$ to generate a recognition network at the phone-level $phone\mathcal{RN}$. Therefore, the next combination is $phone\mathcal{RN}\,o\,\mathcal{L}$ that gives a word-level recognition network by integrating lexical knowledge. After pruning, this word-level network is composed with $\mathcal{G}_{bigram}$ to integrate bigram LM information. A bigram LM was employed first to keep the memory usage within limits. The final grammar-constrained word-level $d\mathcal{RN}$ is thus generated and sent to the trigram LM rescoring module, as shown in the bottom part of Fig. 3 to re-order the decoding paths embedded in $d\mathcal{RN}$. 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.

V. System Setup and Experimental Results

All systems were evaluated on the speaker-independent, 5,000-word WSJ0 (5k-WSJ0) task [18]. The training material consisted of 7,077 utterances from 84 speakers (SI-84), i.e., 15.3 hours of speech. A cross-validation (CV) set was generated by extracting 200 sentences from the SI-84 set. This CV subset accounted for about 3% of the whole training material set and was used to stop the training iterations of the detection-based attribute detectors and event merger. The remaining 6,877 SI-84 sentences were utilized to train the knowledge module. Evaluation was performed on the Nov92 evaluation set, consisting of 330 utterances from 8 speakers.

A. Configuration of Comparison Systems

The proposed system, referred to as DET-WFSM, so called because of detection-based nature of part of the front-end, was described in Section IV-C. In order to demonstrate that our system really benefits from using the WFSM with the particular pruning mechanism and not from applying discriminative ANN-based input features, the output of the front-end shown in Fig. 3 was used to build a *TANDEM* system [27]. Moreover, the performance of two hybrid HMM systems trained on the same SI-84 and evaluated on the same Nov92 test set reported in [29] are also included here for comparison. These systems differ in their feed-forward MLP architectures and are denoted LOQ-1 and LOQ-2, respectively[2]. It is noted that both the *TANDEM* and the hybrid HMM/ANN techniques employed a top-down integrated search approach to ASR.

For the sake of comparison, a baseline HMM/GMM LVCSR system was also built using HTK [7]. The parameters of the acoustic models were estimated using classical ML estimation [54], [55]. Moreover, another HMM/GMM LVCSR system, referred to as MMI-HMM, was built using the MMI criterion [11] with the ML-HMM as seed models. The reader is referred to [26] for additional details about the baseline systems.

A decoupled search approach was also simulated using the standard HMM based acoustic models. Word recognition was accomplished as follows: (i) producing word-level lattices by lexical access on pruned phone-level lattices generated by the MMI-HMM system along with a phone bigram language network trained on phonetic transcriptions of the training part; (ii) composing a pruned version of these word-based lattices by applying a bigram LM; (iii) rescoring the pruned word-lattices with a trigram LM; and (iv) computing the best path. This system is referred to as decoupled MMI-HMM.

Since a stage-by-stage detection is performed at the attribute, phone and word levels we can simultaneously obtain recognition results at all stages. These results are reported in the following two subsections, the first concerning front-end acoustic processing of attribute detection and phone recognition, and the second dealing with LVCSR after back-end lexical and syntactical knowledge integration. Finally in the last two subsections, system combination strategies that achieve the best performance are described in detail.

B. Attribute Detection and Phone Classification Results

The results reported in the second column of Table I are the frame-based accuracies obtained using attribute detectors as described in Section IV-A. The third column of Table I shows frame-based classification accuracy obtained with a Naïve algorithm. This algorithm assigned each frame with the most probable label (true or false). That is, when the majority of the frames in the training set is true for an attribute, then the Naïve algorithm assigns value "true" to that attribute for all frames. This information has been included in Table I to demonstrate that the proposed solution can attain a classification result better than chance.

[2]Note that the full SI-84 material was used to train the LOQ-1 and LOQ-2 systems, which means a 3% additional amount of training data than those used in our proposed WFSM implementation.

TABLE I
CLASSIFICATION ACCURACIES (IN %) AT A FRAME LEVEL FOR THE SPEECH ATTRIBUTES USED IN THIS WORK.

Attribute	Accuracy	Naïve
anterior	93.2	63.8
approximant	95.9	90.8
back	92.9	80.4
continuant	89.9	55.7
coronal	93.1	74.5
dental	99.1	98.9
fricative	95.4	84.7
glottal	99.7	99.2
high	94.9	83.3
labial	92.5	89.0
low	96.9	90.7
mid	93.6	88.2
nasal	97.1	91.3
retroflex	98.4	93.8
round	93.4	85.3
stop	94.9	84.7
tense	90.5	60.5
velar	98.4	94.6
voiced	95.4	59.9
vowel	91.3	67.5

TABLE II
PHONE ACCURACY RATES (PARS) AT A FRAME LEVEL.

Frame-based PAR	Training	CV	Nov92
Detection-based FrontEnd	88.2%	86.2%	85.5%
Phoneme Classifier	87.9%	85.4%	84.3%
FRONT-END in Figure3	89.1%	86.3%	85.8%

Although very good attribute classification accuracies can be obtained for several attributes, such as nasal (97.1%), and stop (94.9%). A deeper analysis revealed that the F1 score, which can be interpreted as a harmonic mean of the precision and recall, is very low for some attributes, such as dental, mid, and velar. For these attributes, the false rejection rate of the attribute class is much higher than that of the corresponding non-attribute class. On the other hand the false acceptance rate of the non-attribute class was instead much higher than that of the attribute class. Thus, most of the test speech was recognized as the non-attribute class. A different training scheme that directly optimizes the performance metrics of interest, for instance, maximal figure-of-merit (MFoM) learning [56] to maximize F1 of the training set, can be used to address that issue.

The classification accuracies at the output of the "FRONT-END" block shown in Fig. 3 are reported in the last row of Table II. The phoneme classification accuracies at the output of the "FRONT-END" block are equal to 89.1%, 86.34%, and 85.82% on the training, cross-validation, and test materials, respectively. These results confirm that high-accuracy phoneme classification with the proposed front-end. Moreover, the closeness between the accuracies on the CV and the evaluation sets indicates that over-fitting has been avoided. For the sake of completeness, the classification results for the "Detection-based FrontEnd," and the "Phoneme Classifier" are separately given in the first and second row of Table II, respectively.

C. Word Recognition Results

Table III shows the performance of all LVCSR systems that employ an ANN to generate the input features (i.e., *TANDEM*). The proposed DET-WFSM system achieved a WER of 6.0%

TABLE III
COMPARING WERS ON THE NOV92 FOR SEVERAL SYSTEMS.

ASR Paradigm	System	WER (%)
Top-down Integrated ASR system	*TANDEM*	6.9
	LOQ-1 [29]	8.4
	LOQ-2 [29]	6.5
Bottom-up Decoupled ASR system	DET-WFSM	6.0

TABLE IV
WER ON THE NOV92 FOR VARIOUS HMM/GMM SYSTEMS.

ASR Paradigm	System	WER (%)
Top-down Integrated ASR system	ML-HMM	5.0
	MMI-HMM	4.6
Bottom-up Decoupled ASR system	DET-WFSM	6.0
	decoupled MMI-HMM	13.3

using the decoupled composition scheme outlined in (3). As shown in Table III, this result represents a relative WER reduction of 28.5%, and 7.7% over LOQ-1, and LOQ-2, respectively. That therefore demonstrates the viability of the proposed approach. Furthermore, since the DET-WFSM system outperforms the *TANDEM* system, which attained a WER of 6.9% as shown in Table III the good word recognition result was not due to the use of discriminative inputs but to the decoding techniques adopted in this study. Finally, if the front-end shown in the upper part of Fig. 3 were to be used within the hybrid HMM/ANN framework, the attainable WER would be 6.5%.

It is noted that the ML-HMM system attained a WER of 5.0%, as shown in the first column of Table IV. With MMI training, the WER can be reduced to 4.6%, as shown in the second row of Table IV. These WERs can be further reduced using advanced discriminative training algorithms (e.g., [13], [57]), and a WER of 3.9% has been reported in [57] when the same experimental working conditions used in this paper are adopted. It should be also mentioned that a WER of 3.0% on the Wall Street Journal task in [58]; nonetheless, a different experimental setup and more complex HMMs have been employed. It is interesting to point out again that the proposed decoupled approach is made possible by the high-accuracy phone classification attained by the proposed front-end. In fact, the decoupled MMI-HMM system delivered a WER as high as 13.3%, as shown in Table IV.

Although it could be argued that a fully integrated decoding scheme is always used in the standard HMM-based system, it is noted that integration of useful information in the speech knowledge hierarchy is often hampered by the integrated approach. On the other hand, the decoupled approach provides a potential solution to LVCSR that can be beneficial to performance and robustness in some specific context where the standard integrated approach is known to fail. For example, for ill-formed utterances, such as those observed in spontaneous speech, partial understanding is often needed because it may not be easy for an integrated approach to properly represent the overall task constraints [36].

D. System Combination Results

A visual inspection of the decoding outcome generated by the decoupled DET-WFSM systems revealed that the target sentence, when it exists, always ranks high among the best candidates. Furthermore, the top-best decoded sentence was often grammatically incorrect, and the trigram LM score for this sentence was lower than that of the grammatically correct sentence. We believe that this kind of recognition error was caused by the strong acoustic models (high-accuracy phone recognition has been reported in this work) that impose a specific recognition path not strictly observing the grammar information which was only integrated at a later stage. To verify that, the standard ML-HMM system was used to re-order the N-best list generated by the decoupled DET-WFSM system. This rescoring scheme was applied to $N = 1000$ generated by the proposed DET-WFSM system, and this reduced the WER from 5% to 4.7% as shown in the second column of Table V, representing a 6% relative WER reduction over the ML-HMM result.

TABLE V
WERS AFTER RESCORING 1000-BEST LISTS GENERATED BY THE DECOUPLED DET-WFSM SYSTEM.

System	ML-HMM	+WFSM-lattices	+DET-scores
WER	5.0%	4.7%	4.3%

Indeed, this result also implies that the proposed approach generates different and potentially better N-best hypotheses than the standard ML-HMM system where a fully-integrated network was used to perform decoding. This outcome is quite important. In fact, it is known that neither determinization nor minimization is performed over the lattices generated by HDecode, and HDecode therefore generates a less accurate set of competing hypotheses than that obtainable by implementing a conventional top-down ASR system with integrated search within the WFSM paradigm [24], [59]. On the other hand, the minimization and determinazation are not performed to the final recognition network generated by (3), as commonly done for ASR system implemented with WFSMs [24], [59]. This outcome is quite interesting, because it confirms that the composition scheme adopted can also provide high-quality lattices, which are of fundamental importance to perform bottom-up ASR with decouple search . We will come back to reconfirm this important assertion in Section V-E when we present rescoring results with the word lattices generated by the HMM systems.

After re-ordering the N-best list, several grammatically incorrect sentences were properly decoded. However, other utterances that would be correctly recognized by the proposed decoupled system might be now incorrectly recognized. We decided to combine the acoustic scores of the ML-HMM system with those of the decoupled DET-WFSM system using the N-best lists generated by the latter system. The acoustic score combination was performed at a sentence-by-sentence level, and 60% of the weight was given to the ML-HMM system scores and 40% to the DET-WFSM scores. By re-ordering the N-best list according to these combined scores and trigram language scores, the WER was further reduced to 4.3%, as shown in the third column of Table V. This result represents a 12% relative improvement over the standard ML-HMM technology and an over 20% relative WER reduction from the proposed decoupled DET-WFSM system with a 6% WER.

It should be pointed out that system combination has been proposed in the past. For example, ROVER [60] performs

TABLE VI
WERs After Rescoring 1000-Best Lists Generated by the Decoupled DET-WFSM System.

System	MMI-HMM	+WFSM-lattices	+DET-scores
WER	4.6%	4.3%	**3.9%**

TABLE VII
WERs After Rescoring 1000-Best Lists Generated by HMM System.

System	Baseline	+HMM-lattices +DET-scores
ML-HMM	5.0%	4.7%
MMI-HMM	4.6%	4.3%

system combination through a majority vote decision. ROVER was not employed in this study because the aim of our rescoring experiment was to shed light upon the quality of the set of top competing hypotheses generated with our DET-WFSM system. High-accuracy hypotheses are needed to carry out bottom-up decouple recognition, and a lack of accurate interfaces between layers doomed the early attempts of building top-down decoupled ASR systems to failure.

A next step is then to replace the ML-HMM system by the discriminatively trained MMI-HMM system in the combination scheme. Furthermore, a WER of **3.9%** was obtained if both DET-WFSM and MMI-HMM acoustic models were taken into account to re-order the N-best lists generated with the proposed technique, as shown in the third column of the first row of Table VI. Again, this acoustic score combination was performed at a sentence-by-sentence level, and 90% of the weight was given to the ML-HMM system scores and 10% to the DET-WFSM scores. The weights were determined using the CV set. This final result represents a relative improvement in performance of 15.1%.

E. Lattice Quality Comparison

Another important point is, as we have highlighted earlier: it can be argued that the lattices generated by HMMs and detectors possess different properties which can benefit system combination. We used an indirect way to compare the quality of the lattices, namely we use only the N-best lists generated by the HMM-based lattices and the DET-WFSM lattices with trigrams rescoring on the 1000-best generated both lattices. Indeed, the WER went from 5.0% to 4.7% for the ML-HMM system, as shown in the second row of Table VII. If an MMI-HMM system was used, the WER went from 4.6% to 4.3%, as shown in the third row of Table VII. When compared with combination results obtained with the DET-WFSM 1000-best lists, we observed a WER reduction from 6.0% to 4.3% and 3.9% with the ML-HMM and MMI-HMM scores, respectively. This seems to confirm that the quality of the word lattices generated with our decoupled DET-WFSM system, although determinazation and minimization have not been applied to the final recognition network generated by (3), is still potentially better than both the lattices generated by the integrated ML-HMM and MMI-HMM systems.

VI. Discussion

For LVCSR tasks, the compiled FSN was often too large and therefore it becomes computationally expensive to find the best

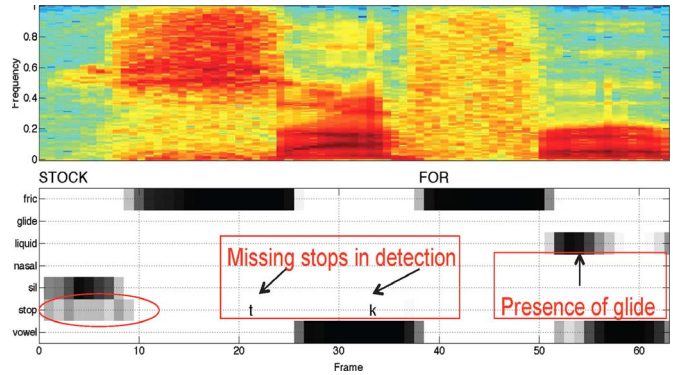

Fig. 5. Spectrogram (upper panel) and posteriogram (lower panel) for the sentence numbered 446c0210 of the Nov92 test set with focus on the area of the errors occur. A conventional LVCSR system miss-recognize the word "Safra," and generates the transcription "Stock for." In the second panel, the time evolution of the posterior probabilities, namely posteriogram, of manner of articulation shows that 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.

sentence through a huge and ever-expanding search space. Thus all knowledge sources must be kept simple in order to be efficiently combined into a single search space. This has particularly inhibited progress at the linguistic level, and almost all current LVCSR systems employ sub-optimal linguistic components such as static lexicons (lexicalization of morphological processes) and n-gram language models (LMs) which force the decoding process to generate hypotheses which sometimes conflict with the acoustic constraints.

Two WSJ examples that were wrongly recognized by the top-down HMM-based systems with integrated search are discussed in the following to illustrate the potential feasibility of using modular search to correct erroneous recognition results. The first is about incorporating manner of articulation information and the second about adding correct suprasegmental information from pitch and duration. It will soon be clear that the incorporation of such low-level events into top-down state-of-the-art HMM systems is often not easily done, but it can be accomplished at a modular level in bottom-up integration. Various knowledge sources can also be integrated in the proposed bottom-up framework by other researchers.

A. Integrating Attribute Information Into Modular Search

For example in the WSJ task, we had observed that a conventional LVCSR 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 Fig. 5 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 at each frame, known as a posteriogram [61] generated by a bank of ANN-based detectors for manner of articulation displayed in the lower panel of Fig. 5, 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

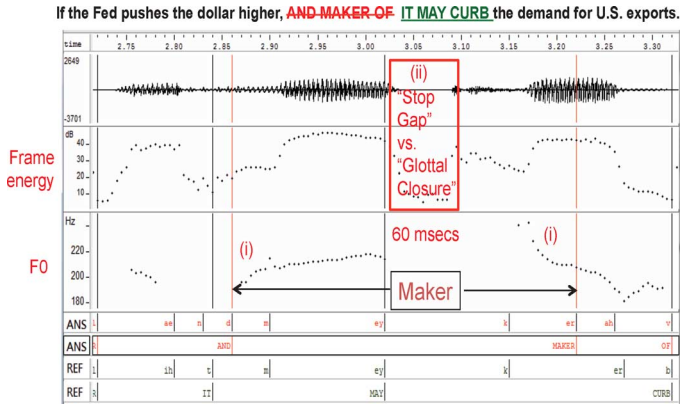

Fig. 6. Prosodic analysis of the WSJ sentence *if the Fed pushes the dollar higher, it may curb the demand for U.S. exports*. The first panel shows the waveform in the time frame between 2.72 and 3.32 seconds, where a recognition error occurs. Specifically *it may curb* is recognized as *and maker of*. The second panel shows the frame energy, whereas the F0 is shown in the third panel. The recognized phone and word sequences are reported in the forth and fifth panel, respectively. The reference phone and word transcriptions are displayed in the sixth and seventh panel, respectively. Two inconsistencies: (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 in "maker" is too long.

information could be properly extracted and included into integrated search, these errors could have been avoided. This misrecognized utterance was corrected with attributed-based lattice rescoring [32], [62] by combining attribute detection scores with the frame HMM log-likelihood scores.

B. Integrating Prosody Information Into Modular Search

In a similar way, supra-segmental information, such as prosody, and long-term language constraints, such as trigram word probabilities, could not be easily cast into the FSN specification when performing top-down knowledge integration. Potential errors could be corrected by using suprasegmental pitch contours and duration features, as demonstrated by a visual inspection of Fig. 6[3]. 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 to show a three-word recognition error occurring when using the same HMM-based system for illustrating the example in Fig. 5 earlier.

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. The reference phone and word transcriptions are displayed in the sixth and seventh panels, respectively. Knowledge-based analysis of 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. In a recent study we found that duration information can be incorporated in a bottom-up manner to improve state-of-the-art HMM system performance. Combining prosody and attribute information was also beneficial. We will report this set of new results in a future paper [63].

[3]Dr. Chen-Yu Chiang of NCTU, Taiwan, had generated this figure.

C. Event Combination Beyond WFSM

In this study only WFSM was used to combine low-level events. New and flexible techniques are needed to accomplish merging of detected events in the probabilistic lattices generated in the intermediate stages during bottom-up knowledge integration. CRFs can be used to describe sequences of symbols (such as phones or words) in terms of input features – e.g., local phonetic attribute detections, and SCARF [46] allows researchers to advance state-of-the-art ASR performance by harnessing large numbers of independent features without having to develop large scale HMM-based LVCSR systems. This will serve as a good tool for bottom-up collaborative ASR research as well. Novel scientific ideas can thus be developed, evaluated and utilized by adding derived acoustic and phonetic attribute features from bottom-up to the existing state-of-the-art LVCSR systems.

VII. Conclusion

We have presented a bottom-up, stage-by-stage, knowledge-integration approach to LVCSR using the AT&T WFSM toolkit. Based on this publicly available set of library routines we were able to extensively evaluate the feasibility of decoupled search strategies with the 5000-word LVCSR WSJ task. Based on our experimental results it was first found that a conventional decoupled HMM-based system cannot deliver the required precision and resulted in a dramatically degraded accuracy. On the other hand, to be competitive with the performance of the state-of-the-art, integrated ASR search strategies, two critical research issues, precision of the output lattice and compactness of the input lattice at every intermediate stage, are need to be further investigated.

Furthermore, the decoupled bottom-up WFSM system gave an overall word error rate of 6% compared with the best WER of 4.6% attained by the HMM-based integrated search system with MMI-based acoustic models. Clearly this was caused by many pruning errors to maintain input lattices at reasonable sizes in order to perform WFSM-based knowledge-integration on memory-limited workstations. Nevertheless we believe the word-level lattices generated with the proposed decoupled WFSM system are more accurate than the ones obtained with the standard HMM beam-search technology. This was demonstrated in our setting where we obtained a 6% relative WER improvement over the standard integrated ML-HMM system by re-ordering 1000-best lists generated with our decoupled approach. A further WER reduction was also observed when the acoustic scores of the HMM and decoupled WFSM systems are combined, yielding a WER of 4.3%. Finally, very good performance for the Nov92 task was attained by combining the acoustic scores of the proposed WFSM system with the acoustic scores of a discriminatively trained MMI-HMM-based system over the top N-best lists embedded in the word lattices generated by the proposed system. The best WER reported in this work is 3.9%, which is comparable with several state-of-the-art modeling techniques applied to the same task [57]. A WER of

3.0% on the Wall Street Journal task in [58] has been recently reported, yet a different experimental setup and more complex HMMs have been used.

A limitation of the detection-based front-end is the lack of temporal overlapping (i.e., asynchrony) characteristics in the attributes across different dimensions. Such asynchrony is central to modern phonological theory (see a review in [64]). Incorporation of asynchrony will significantly modify the attribute targets in running speech in a principled and parsimonious way (e. g., [20]). For spontaneous speech that exhibits significantly more variation in pronunciation than read-style speech, such asynchrony plays a more important role. With the attribute targets modified in a phonologically meaningful manner, we hope that our approach will further enhance the value of the attributes for making word recognition and spontaneous speech recognition more accurate. In addition, detectors based on mechanisms other than feed-forward multi-layer perceptrons, such as deep learning [17], [65], [66], are of great interest as well. In [67], it was demonstrated that high accuracies for both phonological attribute detection and phone estimation accuracy can be attained using DNNs. In [68] was shown that language-independent speech recognition can be performed using phonological attribute detectors; furthermore, phonological attribute have proven useful in automatic language recognition tasks [69], [70].

ACKNOWLEDGMENT

The authors thank Prof. K. Shinoda and his group at Tokyo Institute of Technology for sharing their computational resources during the initial phase of this study. The authors also gratefully acknowledge their colleagues, I. Kim and I-F. Chen, for their assistance during the setup of the large-memory workstations at Georgia Institute of Technology.

REFERENCES

[1] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. PAMI-5, no. 2, pp. 179–190, Mar. 1983.

[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, Feb. 1999.

[3] S. E. Levinson, "Structural methods in automatic speech recognition," *Proc. IEEE*, vol. 73, no. 11, pp. 1625–1650, Nov. 1985.

[4] H. Ney and S. Ortmanns, "Progresses in dynamic programming search for LVCSR," *Proc. IEEE*, vol. 88, no. 8, pp. 1224–1240, Aug. 2000.

[5] J.-L. Gauvain and L. Lamel, "Large vocabulary continuous speech recognition: Advances and applications," *Proc. IEEE*, vol. 88, no. 8, pp. 1181–1200, Aug. 2000.

[6] 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, Aug. 2000.

[7] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book (for HTK Version 3.3)*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[8] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[9] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of the parameters of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, 1995.

[10] D. Zhu, B. Ma, and H. Li, "Speaker verification with feature-space MAPLR parameters," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 505–515, Mar. 2011.

[11] L. R. Bahl, P. F. Brown, P. V. deSouza, and R. L. Mercer, "Maximum mutual information estimation of HMM parameters for speech recognition," in *Proc. ICASSP*, Tokyo, Japan, Apr. 1986, pp. 701–704.

[12] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification minimum error rate methods for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 257–265, May 1997.

[13] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, Orlando, FL, May 2002, pp. 105–108.

[14] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proc. ICASSP*, Philadelphia, PA, Mar. 2005, pp. 961–964.

[15] J. Li, S. Siniscalchi, and C.-H. Lee, "Approximate test risk minimization through soft margin estimation," in *Proc. ICASSP*, Honolulu, HI, Apr. 2007, pp. IV-653–IV-656.

[16] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. ICASSP*, San Francisco, CA, Mar. 1992, pp. 517–520.

[17] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 437–440.

[18] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. ICSLP*, Banff, AB, Canada, Oct. 1992, pp. 899–902.

[19] L. Deng, "Computational models for speech production," in *Computational Models for Speech Pattern Processing*. New York: Springer-Verlag, 1999, NATO ASI.

[20] L. Deng, "Articulatory features and associated production models in statistical speech recognition," in *Computational Models for Speech Pattern Processing*. New York: Springer-Verlag, 1999, NATO ASI.

[21] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech, Audio Proc.*, vol. 2, no. 1, pp. 115–132, Jan. 1994.

[22] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics*, vol. 16, pp. 55–76, 1988.

[23] L. Deng, "Computational models for auditory speech processing," in *Computational Models for Speech Pattern Processing*. New York: Springer-Verlag, 1999, NATO ASI.

[24] M. Mohri, "Finite-state transducers in language and speech processing," *Computat. Linguist.*, vol. 23, no. 2, pp. 269–311, 1997.

[25] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "Openfst: A general and efficient weighted finite-state transducer library," in *Proc. CIAA*, Prague, Czech Republic, Jul. 2007, pp. 11–23.

[26] S. M. Siniscalchi, C.-H. Lee, and T. Svendsen, "A bottom-up stepwise knowledge-integration approach to large vocabulary continuous speech recognition using weighted finite state machines," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 1825–1828.

[27] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, Istanbul, Turkey, Jun. 2000, pp. 1635–1638.

[28] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Boston, MA: Kluwer, 1994.

[29] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. de Mori, "Òlinear hidden transformations for adaptation of hybrid ANN/HMM models," *Speech Commun.*, vol. 49, no. 10-11, pp. 827–835, 2007.

[30] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, "From HMM's to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 360–378, Sep. 1996.

[31] M. Ostendorf, "Moving beyond the Beads-On-A-String model of speech," in *Proc. IEEE ASRU*, Keystone, CO, Dec. 1999, pp. 79–84.

[32] S. M. Siniscalchi and C.-H. Lee, "A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition," *Speech Commun.*, vol. 51, pp. 1139–1153, 2009.

[33] K. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT Press, 1998.

[34] R. Schwartz and Y.-L. Chow, "The N-best algorithm: An efficient and exact procedure for finding N most likely sentence hypotheses," in *Proc. ICASSP*, Albuquerque, NM, Apr. 1990, pp. 81–84.

[35] H. Murveit, J. Butzberger, V. Digalakis, and M. Weintraub, "Large-vocabulary dictation using SRI's DECIPHER(TM) speech recognition system: Progressive-search techniques," in *Proc. ICASSP*, Minneapolis, MN, Apr. 1993, pp. 319–322.

[36] T. Kawahara, C.-H. Lee, and B.-H. Juang, "Flexible speech understanding based on combined key-phrase detection and verification," *IEEE Speech Audio Process.*, vol. 6, no. 5, pp. 558–568, Nov. 1998.

[37] M. D. Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. V. Compernolle, "Template-based continuous speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1377–1390, May 2007.

[38] M. Tang, S. Seneff, and V. W. Zue, "Modeling linguistic features in speech recognition," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 2585–2588.
[39] M. Hasegawa-Johnson *et al.*, "Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop," in *Proc. ICASSP*, Philadelphia, PA, May 2005, pp. 213–216.
[40] K. Demuynck, T. Laureys, D. V. Compernolle, and H. V. Hamme, "FLaVoR: A flexible architecture for LVCSR," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 1973–1976.
[41] K. Demuynck, D. V. Compernolle, and H. V. Hamme, "Robust phone lattice decoding," in *Proc. Interspeech*, Pittsburgh, PA, Sep. 2006, pp. 1622–1625.
[42] J.-T. Chien and C.-H. Chueh, "Joint acoustic and language modeling for speech recognition," *Speech Commun.*, vol. 52, no. 3, pp. 223–235, 2010.
[43] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, Williamstown, MA, Jun./Jul. 2001, pp. 282–289.
[44] G. Zweig, P. Nguyen, D. V. Compernolle, K. Demuynck, L. Atlas, P. Clark, G. Sell, M. Wang, F. Sha, H. H. D. Karakos, A. Jansen, S. Thomas, G. S. V. S. Sivaram, S. Bowman, and J. Kao, "Speech recognition with segmental conditional random fields: A summary of the JHU CLSP 2010 summer workshop," in *Proc. ICASSP*, Prague, Czech Republic, May 2011, pp. 5044–5047.
[45] G. Zweig, "Classification and recognition with direct segment models," in *Proc. ICASSP*, Kyoto, Japan, Mar. 2012, pp. 4161–4164.
[46] G. Zweig and P. Nguyen, "SCARF: A segmental conditional random field toolkit for speech recognition," in *Proc. Interspeech*, Makuhari, Japan, Sep. 2010, pp. 2858–2861.
[47] 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.
[48] I. Bromberg, Q. Fu, J. Hou, J. Li, C. Ma, B. Matthews, A. Moreno-Daniel, J. Morris, S. M. Siniscalchi, Y. Tsao, and Y. Wang, "Detection-based ASR in the automatic speech attribute transcription project," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 1829–1832.
[49] H. Hermansky and S. Sharma, "Temporal patterns (TRAPS) in ASR of noisy speech," in *Proc. ICASSP*, Phoenix, AZ, Mar. 1999, pp. 289–292.
[50] P. Schwarz, P. Matějaka, and J. Černockýý, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 325–328.
[51] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "Towards bottom-up continuous phone recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. and Understanding*, Kyoto, Japan, Dec. 2007, pp. 566–569.
[52] H. Ketabdar and H. Bourlard, "Hierarchical integration of phonetic and lexical knowledge in phone posterior estimation," in *Proc. ICASSP*, Las Vegas, NV, Mar. 2008, pp. 4065–4068.
[53] S. Haykin, *Neural Networks: A Comprehensive Foundation*. London, U.K.: Macmillan, 1994.
[54] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, vol. 41, pp. 164–171, 1970.
[55] L. R. Liporace, "Maximum likelihood estimation for multivariate observation of Markov sources," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 5, pp. 729–734, Sep. 1982.
[56] S. Gao, W. Wu, C.-H. Lee, and T.-S. Chua, "A maximal figure-of-merit (MFoM)-learning approach to robust classifier design for text categorization," *ACM Trans. Inf. Syst.*, vol. 24, no. 2, pp. 190–218, 2006.
[57] J. Li, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Georgia Inst. of Technol., Atlanta, GA, 2008.
[58] G. Heigold, S. Wiesler, M. Nussbaum, P. Lehnen, R. Schlüter, and H. Ney, "Discriminative HMMs, Log-Linear Models, and CRFs: What is the difference?," in *Proc. ICASSP*, Dallas, TX, Mar. 2010, pp. 5546–5549.
[59] M. Mohri, F. C. N. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Comput. Speech Lang.*, vol. 16, pp. 69–88, 2002.
[60] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (rover)," in *Proc. IEEE ASRU Wkshp*, Santa Barbara, CA, Dec. 1997, pp. 347–352.
[61] P. Fousek and H. Hermansky, "Towards ASR based on hierarchical posterior-based keyword recognition," in *Proc. ICASSP*, Toulose, France, May 2006, pp. 433–436.
[62] S. M. Siniscalchi, P. Schwarz, and C.-H. Lee, "High-accuracy phone recognition by combining high-performance lattice generation and knowledge based rescoring," in *Proc. ICASSP*, Honolulu, HI, Apr. 2007, pp. IV–869–V–872.
[63] C.-Y. Chiang, S. M. Siniscalchi, S.-H. Chen, and C.-H. Lee, "A study on cross-language knowledge integration in mandarin LVCSR," in *Proc. 8th Int. Symp. Chinese Spoken Lang. Process. (ISCSLP-12)*, Dec. 2012, pp. 315–319.
[64] L. Deng and D. O'Shaughnessy, *Speech Processing: A Dynamic and Optimization-Oriented Approach (Signal Processing and Communications)*. New York: Marcel Dekker, 2003.
[65] A. R. Mohamed, G. Dahl, and G. E. Hinton, "Deep belief networks for phone recognition," in *Proc. NIPS 22 Workshop Deep Learning for Speech Recognit. Related Applicat.*, Whistler, BC, Canada, Dec. 2009.
[66] D. Yu and L. Deng, "Deep convex network: A scalable architecture for deep learning," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 2285–2288.
[67] D. Yu, S. M. Siniscalchi, L. Deng, and C.-H. Lee, "Boosting attribute and phone estimation accuracy with deep neural networks for detection-based speech recognition," in *Proc. ICASSP*, Kyoto, Japan, Mar. 2012, pp. 4169–4172.
[68] 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, Lang. Process.*, vol. 20, no. 3, pp. 875–887, Mar. 2012.
[69] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Exploring universal attribute characterization of spoken languages for spoken language recognition," in *Proc. Interspeech*, Brighton, U.K., Sep. 2009, pp. 168–171.
[70] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 209–227, 2013.

Sabato Marco Siniscalchi (M'07) is an Assistant Professor at the University of Enna "Kore," and an Affiliate with the Georgia Institute of Technology. He received his Laurea and Doctorate degrees in Computer Engineering from the University of Palermo, Palermo, Italy, in 2001 and 2006, respectively. In 2001, he was employed by STMicroelectronics where he designed optimization algorithms for processing digital image sequences on very long instruction word (VLIW) architectures. In 2002, he was adjunct professor at the University of Palermo and taught several undergraduate courses for Computer and Telecommunication engineering. In 2006, he was a Post Doctoral Fellow at the Center for Signal and Image Processing (CSIP), Georgia Institute of Technology, Atlanta, under the guidance of Prof. C.-H. Lee. From 2007 to 2009, he joined the Norwegian University of Science and Technology, Trondheim, Norway, as a Research Scientist at the Department of Electronics and Telecommunications under the guidance of Prof. T. Svendsen. In 2011 he became a Georgia Tech affiliate. His main research interests are in speech processing, in particular automatic speech and speaker recognition, and language identification.

Torbjørn Svendsen (M'85–SM'11) is a professor at the Department of Electronics and Telecommunications at the Norwegian University of Science and Technology (NTNU). Dr. Svendsen received the siv.ing (MSc) and dr.ing. degrees from the Norwegian Institute of Technology (NTH) in 1980 and 1985, respectively.

Dr. Svendsen has been a research scientist at SINTEF before joining NTH as an associate professor in 1988. Since 1995 he has been professor of speech processing at NTNU. He has had extended research stays at AT&T Bell Laboratories, Murray Hill, New Jersey; AT&T Labs, Florham Park, New Jersey; Griffith University, Brisbane, Australia and Queensland University of Technology, Brisbane, Australia. His research interests include automatic speech recognition; speech synthesis; speech coding and speech analysis and modeling. He has authored and co-authored more than 70 papers in these areas.

Prof. Svendsen is a member of the IEEE Signal Processing Society (SPS) and the International Speech Communication Association (ISCA). He has been a member of the IEEE SPS Speech Processing Technical Committee.

Chin-Hui Lee (S'78–M'81–SM'91–F'97) is a professor at the School of Electrical and Computer Engineering at the Georgia Institute of Technology. Dr. Lee received the B.S. degree in Electrical Engineering from National Taiwan University, Taipei, in 1973, the M.S. degree in Engineering and Applied Science from Yale University, New Haven, in 1977, and the Ph.D. degree in Electrical Engineering with a minor in Statistics from University of Washington, Seattle, in 1981.

Dr. Lee started his professional career at Verbex Corporation, Bedford, MA, and was involved in research on connected word recognition. In 1984, he became affiliated with Digital Sound Corporation, Santa Barbara, where he engaged in research and product development in speech coding, speech synthesis, speech recognition and signal processing for the development of the DSC-2000 Voice Server. Between 1986 and 2001, he was with Bell Laboratories, Murray Hill, New Jersey, where he became a Distinguished Member of Technical Staff and Director of the Dialogue Systems Research Department. His research interests include multimedia communication, multimedia signal and information processing, speech and speaker recognition, speech and language modeling, spoken dialogue processing, adaptive and discriminative learning, biometric authentication, and information retrieval. From August 2001 to August 2002 he was a visiting professor at School of Computing, The National University of Singapore. In September 2002, he joined Georgia Institute of Technology.

Prof. Lee has participated actively in professional societies. He is a member of the IEEE Signal Processing Society (SPS), and the International Speech Communication Association (ISCA). In 1991–1995, he was an associate editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. During the same period, he served as a member of the ARPA Spoken Language Coordination Committee. In 1995–1998 he was a member of the Speech Processing Technical Committee and later became the chairman from 1997 to 1998. In 1996, he helped promote the SPS Multimedia Signal Processing Technical Committee in which he is a founding member.

Dr. Lee is a Fellow of the IEEE, and has published about 400 papers and 30 patents. He received the SPS Senior Award in 1994 and the SPS Best Paper Award in 1997 and 1999, respectively. In 1997, he was awarded the prestigious Bell Labs President's Gold Award for his contributions to the Lucent Speech Processing Solutions product. Dr. Lee often gives seminal lectures to a wide international audience. In 2000, he was named one of the six Distinguished Lecturers by the IEEE Signal Processing Society. He was also named one of the two ISCA's inaugural Distinguished Lecturers in 2007–2008. Recently he won the SPS's 2006 Technical Achievement Award for "exceptional contributions to the field of automatic speech recognition". He was a plenary speaker at ICASSP2012 in Kyoto, Japan. More recently he was awarded the 2012 ISCA Medal of Scientific Achievement for "pioneering and seminal contributions to the principles and practices of automatic speech and speaker recognition, including fundamental innovations in adaptive learning, discriminative training and utterance verification".