Leveraging Speech Production Knowledge for Improved Speech Recognition

Abhijeet Sangwan, John H.L. Hansen

Center for Robust Speech Systems (CRSS)  
University of Texas at Dallas (UTD), Richardson, Texas, U.S.A  
{abhijeet.sangwan, John.Hansen}@utdallas.edu

Abstract—This study presents a novel phonological methodology for speech recognition based on phonological features (PFs) which leverages the relationship between speech phonology and phonetics. In particular, the proposed scheme estimates the likelihood of observing speech phonology given an associative lexicon. In this manner, the scheme is capable of choosing the most likely hypothesis (word candidate) among a group of competing alternative hypotheses. The framework employs the Maximum Entropy (ME) model to learn the relationship between phonetics and phonology. Subsequently, we extend the ME model to a ME-HMM (maximum entropy-hidden Markov model) which captures the speech production and linguistic relationship between phonology and words. The proposed ME-HMM model is applied to the task of re-processing N-best lists where an absolute WRA (word recognition rate) increase of 1.7%, 1.9% and 1% are reported for TIMIT, NTIMIT, and the SPINE speech in noise corpora (15.5% and 22.5% relative reduction in word error rate for TIMIT and NTIMIT).

I. INTRODUCTION

In this study, we propose a disambiguation scheme based on phonological features (PFs) that exploits speech production knowledge embedded in phonology. Our motivation stems from research which shows that PFs capture production variability in greater detail than phones [1], [2], [3], [4]. Herein, the inability of phones to effectively model production variability is exposed in the errors made by standard ASR. As shown in Fig. I, production related difficulty is often presented as ambiguous output in standard ASR structures such as N-best lists, lattices or a word-mesh. In these structures, ASR performance can be improved by selecting the correct alternative among ambiguous words. In view of this observation, the proposed scheme is designed to exploit phonological knowledge in order to consistently select correct words from ambiguous alternatives. This paradigm is shown in Fig. I, where the proposed scheme [Block (3)] first extracts PF sequences from the speech signal [Block (2a)], and then uses the PF information to disambiguate confusable words in the standard ASR output [Block (2b)]. In particular, the proposed scheme computes and assigns a “phonological score” to each and every ambiguous word based on the likelihood of jointly observing that word and the corresponding PFs. It is the leveraging of these two domains that makes the proposed solution unique. Thereafter, the new scores are used in conjunction with language model (LM) scores to re-rank the alternate hypotheses, and choose the best candidate. In summary, our contribution is the proposed PF-based framework that allows an efficient and meaningful integration of phonology knowledge into standard ASR systems. The proposed system can always be combined with existing and continuously improving back-end solutions, thereby leveraging the orthogonal knowledge of phonology to solve a common problem of improving ASR performance.

The ability to compute phonological scores for words (or corresponding phone-sequences) lies at the heart of the proposed scheme. In this study, we design and develop a ME-HMM (maximum entropy-hidden Markov model) to learn the probabilistic relationship between phones and their corresponding spectro-temporal articulation variability (via PFs). Subsequently, we apply this newly acquired knowledge to score the joint word-phonology observations. Specifically, the MEM (maximum entropy model) models the phonological variability of phones, and the HMM structure models the temporal evolution of words from a succession of its component phones. Furthermore, a robust contextually-aware input feature-set is also proposed for the MEM. The proposed feature-set uses the knowledge of phonetic-context and co-phonological states to build accurate models of phones-PFs relationship. Finally, the proposed ME-HMM based phonological disambiguation scheme is used to process ASR N-best lists in a speech recognition experiment.

II. GOVERNMENT PHONOLOGY SYSTEM

In this section, we briefly review the basics of Government Phonology (GP) theory employed within the scope of this study.

A. Government Phonology

The GP theory is built on a small set of primes (articulation properties), and rules that govern their interactions [3]. A GP prime produces phones in speech by operating in isolation or in combination with other primes. The primes are broadly categorized into three groups, namely, resonance, manner and source primes. In general, resonance primes govern vowels, manner primes govern articulation of consonants, and source

<table>
<thead>
<tr>
<th>Property</th>
<th>Attributes</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primes</td>
<td>Resonance</td>
<td>A, I, E, U</td>
</tr>
<tr>
<td></td>
<td>Manner</td>
<td>S, h, N</td>
</tr>
<tr>
<td></td>
<td>Source</td>
<td>H</td>
</tr>
<tr>
<td>Headedness</td>
<td>a, i, u</td>
<td>A=a[l]/a+u=a+i=a+u+i/a+y/</td>
</tr>
</tbody>
</table>
primed dictate voicing in speech. Table I lists the GP primes along with examples of how primes generate phonemes.

B. Integrating GP in the Proposed Scheme

The extraction of GP features is performed using an HMM-based scheme described in [1]. Using the GP recognition scheme described above, 11 binary numbers are obtained for each speech frame in the utterance corresponding to the 11 GP elements. The binary representation stems from the fact that each GP element can either be off (0) or on (1) for each speech frame. It is useful to note that since each GP type is decoded independently, the system allows for asynchronous behavior in terms of on/off switching action of the GP elements (e.g., if an unvoiced stop follows a nasal phone, then nasalization and voicing are correctly allowed to turn off at different times). In this manner, $2^{11} = 2048$ unique combinations of GP elements per speech frame are possible. Here, a small subset of combinations represent the canonical form of the phones, while a majority of combinations represent the phonetic-variants. In this study, the canonical GP forms of the phones are taken from [3].

III. PROPOSED ME-HMM FRAMEWORK

In this section, we develop the proposed ME-HMM framework for our phonology-based disambiguation scheme. In the proposed scheme, the ME-HMM models the relationship between a GP-sequence and a phone-sequence. This process is illustrated in Fig. 2, where the ME-HMMs corresponding for words “for” and “force” are shown. The phone sequences for words “for” (/[f]/,[o]/r/) and “force” (/[f]/,[o]/r/,/s/) are shown on the Y-axes, and the corresponding $i^{th}$ GP-sequence is shown on the X-axes. Due to the ambiguity faced by standard
ASR, the $j^{th}$ GP-sequence could have been generated by either “for” or “force” (see 2b in Fig. I). In order to resolve such ambiguity, the proposed ME-HMM computes the likelihood of a word (or equivalent phone-sequence) generating the GP sequence. In this manner, 11 likelihoods, a “phonological score” is obtained for each word. As a result of employing ME-HMMs, 3 scores are now assigned to each ambiguous word in the standard ASR output: acoustic score, language score (by the standard ASR), and phonological score (by the proposed ME-HMM system). Using the newly generated phonological scores, previously assigned to each ambiguous word in the standard ASR output: acoustic score, language score (by the standard ASR), and phonological score (by the proposed ME-HMM system). Using the newly generated phonological scores, previously confusable words can now reassessed and therefore resolved (e.g., “force” vs. “for”). In this manner, the phonological score offers a new dimension of separability among otherwise ambiguous words by leveraging the unique knowledge of speech production.

A. Development of ME-HMM

Let $w$ be a candidate word, and $w \equiv \{p_1, p_2, \ldots, p_M\}$ be the canonical phone-sequence. In other words, $w$ is composed of $M$ phones $p_1, p_2, \ldots, p_M$ in this order. In Fig. 2, this is illustrated for an example word “force” composed of phones $p_1=/f/, p_2=/ao/, p_3=/t/,$ and $p_4=/s/.$ Let the word $w$ span over $N$ speech-frames, where the $j^{th}$ speech frame is denoted by $f_j.$ Furthermore, let the binary state of the $i^{th}$ GP-element ($i = 1, \ldots, 11$) and $j^{th}$ speech frame for word $w$ be given by $g_{ij}.$ The above-defined variables are shown in Fig. 2, where phones are in circles, and GP-states are in squares.

Within the production of $w,$ the exact temporal evolution of phones $p_1, p_M$ is unknown due to the inherent uncertainty in articulation. This uncertainty is captured within the phone-trellis where each path in the trellis is one possible articulation of $w.$ In Fig. 2, one choice of articulation is shown by the path in bold. Furthermore, each path in the trellis determines the phones responsible for the observed GP states $(g_{ij}, j = 1, \ldots, N).$ For example, for the bold path shown in Fig. 2: $g_{i-2} /f/ = 1$ is generated by $/f/,$ $g_{i-1} /ao/ = 1$ is generated by $/ao/,$ and so on. This generative relationship between phones and corresponding GP states is captured by the MEM in the proposed scheme, which is developed in greater detail in Sec. III-B. Next, if all articulation possibilities of $w$ are considered, the forward algorithm [5] can be used to compute the total likelihood of $w$ generating the observed $i^{th}$ GP sequence. In order to use the forward algorithm, the various elements of the HMM framework are first enumerated below:

1) Number of HMM states are equal to the number of canonical phones in $w$ (i.e., $M$).
2) Number of distinct observation symbols for the $i^{th}$ GP sequence are equal to $2,$ (0 or 1)
3) The state transition matrix given by $A$ is defined below,
4) The observation symbol probability distribution given by $b$ is modeled by the MEM, and is developed in Sec. III-B, and,

\[
A = \begin{bmatrix}
\beta & 1 - \beta & 0 & \cdots & 0 \\
0 & \beta & 1 - \beta & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1
\end{bmatrix},
\]

where $a_{mk}$ is the $(m, k)^{th}$ element of $A.$ Furthermore, based on the above arguments, the initial state vector is given by,

\[
\Pi = [1 \ 0 \ \cdots \ 0]^T.
\]

Finally, let the likelihood that the $w$ generated the $i^{th}$ GP sequence be given by $\Lambda_i(w),$ which is computed by the forward algorithm as follows,

\[
\text{Initialization :} \quad \alpha_1(k) = \Pi(k) b(g_{i1}, p_k), k = 1, \ldots, M \ (2)
\]

\[
\text{Recursion :} \quad \alpha_{j+1}(k) = b(g_{ij+1}, p_k) \sum_{m=1}^{M} \alpha_j(m) a_{mk}, \quad k = 1, \ldots, M \ (3)
\]

\[
\text{Termination :} \quad \Lambda_i(w) = \sum_{m=1}^{M} \alpha_M(m), \quad (4)
\]

where $\alpha$ is the likelihood of partial observations. The “phonological-score” $\Lambda_{PF}$ for the word $w$ can now be obtained by summing the likelihoods over all individual GP sequences,

\[
\Lambda_{PF}(w) = \sum_{i=1}^{11} \Lambda_i(w). \quad (5)
\]

The phonological-score $\Lambda_{PF}$ may also be viewed as a production score for the word $w,$ where a high score requires production characteristics of $w$ in the GP space to conform with observed statistical variability. Herein, it is important to note that the lexical output from standard ASR and GP sequence from the GP extraction system are two manifestations of the same underlying speech event. Therefore, it is intuitive to believe that in a list of words, the correct word choice is most likely to show maximum agreement with the corresponding phonology. The proposed ME-HMM system exploits this intuition by leveraging PFs to comprehensively model observed variability in phones. This newly obtained statistical knowledge is then used to score the claim of each word for the list of alternatives.
In the next section, we develop the MEM which serves the role of computing observation symbol generating probabilities $b$ in the above-described HMM structure.

B. Maximum Entropy Modeling

Maximum Entropy (ME) modeling is an evidence based discrete modeling technique. It has been successfully employed in many speech and language tasks such as part-of-speech (POS) tagging [6], machine translation (MT) [7], and acoustic modeling [8]. The ME model (MEM) is extremely flexible since it can relate a variable number of events to the observation. It provides an attractive methodology for relating phones, phonetic context, and high-level information with the GP state of a given frame.

The objective of the MEM is to model the output symbol generation probability distribution ($b$) conditioned on various levels of knowledge. Particularly, the MEM views the different levels of knowledge as evidence. Upon training, the MEM is able to learn weight parameters that quantify the relative importance of different evidences. Within our MEM framework, evidence is implemented as features. Specifically, the feature value is 1 if the evidence is present, and 0 if not.

The different evidence-types are shown in Fig. 3 and listed below:

- (I) Co-occurring phone $p(j)$: For a GP state $g_{ij}$, the phone that occupies that $j^{th}$ frame in the phone-trellis is the co-occurring phone. In Fig. 2, for word “force” /r/ is the co-occurring phone for $g_{ij}$.
- (II) Preceding phone: The preceding phone occurs before the co-occurring phone in the phone-sequence. For example, /ao/ is the preceding phone for the co-occurring phone /r/ in the previous case.
- (III) Succeeding phone: The succeeding phone occurs after the co-occurring phone in the phone-sequence. For example, /s/ is the succeeding phone for the co-occurring phone /r/.
- (IV) Static Co-phonology (Active): The complementary active GP states ($g_{ij}^* = 1$, $i' \neq i$) for the same $j^{th}$ frame. In Fig. 3, for frame $f_j$, the co-phonological states are $S,H,A$ for GP state of $I$.
- (V) Static Co-phonology (Inactive): The complementary inactive GP states ($g_{ij}^* = 0$, $i' \neq i$) for the same $j^{th}$ frame. In Fig. 3, for frame $f_j$, the co-phonological states are $E,U,h,N,a,i,u$ for GP state of $I$.
- (VI) Dynamic Co-phonology (“Steady”): The dynamic knowledge in co-phonological states captures the knowledge of recent state-transitions (0-to-1 or 1-to-0). As shown in Fig. 3, the state-transition of interest are in the neighborhood of 2 frames of the $j^{th}$ frame ($f_{j-2}$ to $f_{j+2}$). If no state-transitions are in this neighborhood, then the co-phonological sequence is in the “steady” state (e.g.,
EVIDENCES

(VII) Dynamic Co-phonology (“Transient”): If state-transitions are in this neighborhood, then the co-phonological sequence is in the “transient” state (e.g., A,S,H in Fig. 3).

Evidences (I), (II), and (III) represent the triphone-context. Furthermore, evidences (IV) and (V) represent the static co-phonological information. Finally, evidences (VI) and (VII) represent the dynamic knowledge in co-phonology. It is noted that moving from evidence types (I) through (VII) constitutes a growing body of evidence. It is expected that MEM quality would improve as newer evidence is incorporated into the modeling.

Finally, we formalize the development of the MEM. Let $E_j$ be the set of evidence available at the $j^{th}$ frame, and $e_l \in E_j$ be the $l^{th}$ evidence as discussed above. Furthermore, let the ME feature for the $i^{th}$ GP element be given by $\mu_i$. Each ME feature is a binary operator on the evidence $e_l$, (i.e., if the feature is observed, it produces a value of unity, otherwise it is zero),

$$\mu_i(e_l) = \begin{cases} 1 & \text{if evidence } e_l \text{ is present,} \\ 0 & \text{otherwise.} \end{cases}$$

For example, in the case of evidence (I): $\mu_i(e_1) = 1$ for co-occurring phone only and $\mu_i(e_1) = 0$ for all other phones.

Now, the MEM can be used to compute the observation symbol probabilities (b) in the ME-HMM model (see Sec. III). In particular, the observation symbol generation computes the probability of observing GP state $g_{ij}$ given the evidence set $E_j$. Let $b_{ij}$ be the likelihood of observing $g_{ij}$ given $E_j$, given by:

$$b_{ij} = p(g_{ij}|E_j) = \frac{1}{Z_\lambda(E)} \exp \left( \sum_{i=1}^{L} \lambda_{il}\mu_i(e_l) \right), \quad (6)$$

where $Z_\lambda(E)$ is a normalization term, and $\lambda_{il}$ are the weights assigned to the ME feature. As mentioned earlier, the weights correspond to the importance of a feature in estimating the likelihood in question. The MEM parameters are learned offline during the training phase. In this study, the ME models were trained using the “Maximum Entropy Modeling Toolkit” [9].

IV. RESULTS AND DISCUSSION

The proposed PF-based disambiguation scheme is applied to the task of re-processing N-Best lists. Particularly, we work with 20-Best lists generated from the test sections of the TIMIT, NTIMIT, and SPINE corpora. As shown in Table II the baseline WEERs (word error rates) obtained for the TIMIT, NTIMIT, and SPINE corpora are 8%, 12.1%, and 37.8% respectively.

Fig. 3. MEM models the conditional probability of observing the GP state $g_{ij}$ given the various phonetic and co-phonological evidence. The (1) co-occurring phone, (2) previous phone, and (3) succeeding phone constitute phonetic evidence. The (4) co-phonological OFF states, (5) co-phonological ON states, (6) co-phonological “transient” states, and (7) co-phonological “steady” states form the co-phonological evidence.
In order to obtain the state-sequences for all GP elements, a separate HMM-based GP extraction scheme was trained for TIMIT, NTIMIT, and SPINE. Furthermore, a separate MEM was also trained for TIMIT, NTIMIT, and SPINE using data from the train-sets of the respective corpora. During test, the HMM-based GP extraction scheme was used to generate the GP state-sequences for the test-sets from TIMIT, NTIMIT, and SPINE. As a result, we obtain the GP state-sequences corresponding to the same utterances for which the 20-Best lists were generated using the above baseline ASR system. Next, each 20-Best list was processed as follows: First the word-level time-segmentation, acoustic-score and language-score for each ambiguous word within the N-best list was identified. As shown in Fig. I, the ambiguous words in an N-best list are readily identified. Using the word-level timing information, the corresponding 11 GP state-sequences for each ambiguous word was identified. Subsequently, by using the proposed ME-HMM model and the forward algorithm described in Sec. III, a “phonological score” for each ambiguous word was computed. Furthermore, the phonological score for each sentence in the 20-Best list was computed by taking a sum of the constituent ambiguous word phonological-scores. Finally, the 20-Best list was re-ranked by using a total hypothesis score which was a simple linear combination of the phonological and language scores. As shown in Table II, the proposed ME-HMM based PF disambiguation scheme achieved an absolute WRA (word recognition accuracy) improvement of 1.7%, 1.9%, and 1% for TIMIT, NTIMIT, and SPINE, respectively as a result of the 20-Best list re-ranking process.

In order to illustrate the nature of improvement obtained by the proposed scheme, we show the part-of-speech (POS) tags for the correctly detected words from the baseline ASR and proposed ME-HMM scheme in Table III. This split-analysis of word recognition serves to illustrate the nature of the improvement obtained by the proposed scheme. The POS tags for the analysis are obtained by means of the tree-tagger tool [10]. From the table, it is observed that the proposed system improves word recognition in all POS categories (adverbs, adjectives, nouns, verbs and others) for both TIMIT and NTIMIT. For the case of SPINE, the proposed scheme achieves improved noun and verb word recognition rates. Herein, the gain in word recognition rates of nouns and verbs is significant, since they tend to be more information-bearing within the utterances, and of particular importance for applications like spoken document retrieval (SDR) [11].

V. Conclusion

In this paper, a novel methodology for speech recognition disambiguation based on the ME-HMM framework was proposed. The proposed ME-HMM framework served to exploit the relationship between low-level signal phonology and higher-level speech phonetics. Subsequently, the ME model was adapted into an HMM framework to form a ME-HMM system which was employed as a tool to compute the likelihood of observing speech segments conditioned on phonological knowledge. In our experiments, words were chosen as the logical speech segments, but the system is just as easily applied to supra- or sub-lexical structures. By computing phonological scores of N-best lists, we were able to resolve ambiguity by achieving a relative WER reduction of 22.5% and 15.7% in the TIMIT and NTIMIT corpora.

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>TIMIT, NTIMIT, AND SPINE: SPEECH RECOGNITION PERFORMANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sub</td>
</tr>
<tr>
<td>----------</td>
<td>-----</td>
</tr>
<tr>
<td>TIMIT</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>6.6</td>
</tr>
<tr>
<td>ME-HMM</td>
<td>4.7</td>
</tr>
<tr>
<td>NTIMIT</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>10.0</td>
</tr>
<tr>
<td>ME-HMM</td>
<td>7.6</td>
</tr>
<tr>
<td>SPINE</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>28.1</td>
</tr>
<tr>
<td>ME-HMM</td>
<td>28.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE III</th>
<th>TIMIT, NTIMIT, AND SPINE: WORD RECOGNITION ACCURACIES FOR DIFFERENT PART-OF-SPEECH ELEMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adjective</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>TIMIT</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>94.7</td>
</tr>
<tr>
<td>ME-HMM</td>
<td>94.7</td>
</tr>
<tr>
<td>NTIMIT</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>92.9</td>
</tr>
<tr>
<td>ME-HMM</td>
<td>93.6</td>
</tr>
<tr>
<td>SPINE</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>62.7</td>
</tr>
<tr>
<td>ME-HMM</td>
<td>62.1</td>
</tr>
</tbody>
</table>

In order to illustrate the nature of improvement obtained by the proposed scheme, a separate HMM-based GP extraction scheme was trained for TIMIT, NTIMIT, and SPINE. Furthermore, a separate MEM was also trained for TIMIT, NTIMIT, and SPINE using data from the train-sets of the respective corpora. During test, the HMM-based GP extraction scheme was used to generate the GP state-sequences for the test-sets from TIMIT, NTIMIT, and SPINE. As a result, we obtain the GP state-sequences corresponding to the same utterances for which the 20-Best lists were generated using the above baseline ASR system. Next, each 20-Best list was processed as follows: First the word-level time-segmentation, acoustic-score and language-score for each ambiguous word within the N-best list was identified. As shown in Fig. I, the ambiguous words in an N-best list are readily identified. Using the word-level timing information, the corresponding 11 GP state-sequences for each ambiguous word was identified. Subsequently, by using the proposed ME-HMM model and the forward algorithm described in Sec. III, a “phonological score” for each ambiguous word was computed. Furthermore, the phonological score for each sentence in the 20-Best list was computed by taking a sum of the constituent ambiguous word phonological-scores. Finally, the 20-Best list was re-ranked by using a total hypothesis score which was a simple linear combination of the phonological and language scores. As shown in Table II, the proposed ME-HMM based PF disambiguation scheme achieved an absolute WRA (word recognition accuracy) improvement of 1.7%, 1.9%, and 1% for TIMIT, NTIMIT, and SPINE, respectively as a result of the 20-Best list re-ranking process.

In order to illustrate the nature of improvement obtained by the proposed scheme, we show the part-of-speech (POS) tags for the correctly detected words from the baseline ASR and proposed ME-HMM scheme in Table III. This split-analysis of word recognition serves to illustrate the nature of the improvement obtained by the proposed scheme. The POS tags for the analysis are obtained by means of the tree-tagger tool [10]. From the table, it is observed that the proposed system improves word recognition in all POS categories (adverbs, adjectives, nouns, verbs and others) for both TIMIT and NTIMIT. For the case of SPINE, the proposed scheme achieves improved noun and verb word recognition rates. Herein, the gain in word recognition rates of nouns and verbs is significant, since they tend to be more information-bearing within the utterances, and of particular importance for applications like spoken document retrieval (SDR) [11].

V. Conclusion

In this paper, a novel methodology for speech recognition disambiguation based on the ME-HMM framework was proposed. The proposed ME-HMM framework served to exploit the relationship between low-level signal phonology and higher-level speech phonetics. Subsequently, the ME model was adapted into an HMM framework to form a ME-HMM system which was employed as a tool to compute the likelihood of observing speech segments conditioned on phonological knowledge. In our experiments, words were chosen as the logical speech segments, but the system is just as easily applied to supra- or sub-lexical structures. By computing phonological scores of N-best lists, we were able to resolve ambiguity by achieving a relative WER reduction of 22.5% and 15.7% in the TIMIT and NTIMIT corpora.

References