IMPROVED SPEECH MODELING AND RECOGNITION USING MULTI-DIMENSIONAL ARTICULATORY STATES AS PRIMITIVE SPEECH UNITS

L. Deng, J. Wu, and H. Sameti

Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

ABSTRACT

In this paper we provide a formal description of a speech recognizer designed on the basis of elaborate articulatory timing that is asynchronous across the multiple articulatory-feature dimensions. Three recently improved critical components of the recognizer are described in detail. Evaluation results, obtained from a standard TIMIT phonetic recognition task confined within the N-best rescoring scenario, are reported on comparative performances between the new feature-based recognizer and a recognizer using the conventional context-dependent triphone units. The results demonstrate an overall superior quality of the rescored N-best list from the feature-based recognizer over that from the triphone-based recognizer. Greater performance improvements are observed as the top number of candidate sentences increases.

1. INTRODUCTION

We have in the past several years pursued the development of an articulatory feature based statistical framework and of the related sub-phonemic units of speech for new speech recognizer design [6, 5, 3]. A main objective of the development is to devise a parsimonious and parametric way for modeling context-dependent behaviors in fluent speech. One unique attribute of our new speech recognizer is its exquisite and elaborate construction of a set of primitive speech units at the feature (subphonemic) level. These units were constructed from multi-dimensional articulatory features overlapping across varying dimensions. Motivated by the theory of distinctive features and by the principles from the more recent articulatory phonology [1], our recognizer demonstrated effectiveness in standard phonetic classification and recognition tasks (TIMIT).

In all the previously published evaluation experiments, the benchmark speech recognizer for comparison purposes was a conventional HMM system using phonemes as the primitive speech units. Context independent nature of the phonemic units has made the performance comparison with the feature-based system subject to criticism in that the power of the latter system comes from its (elegant) exploitation of contextual information. The main purpose of the paper is to report our recent results on performance comparison between the feature-based system and a system based on the conventional triphone units which represent context dependency spanning over several phonetic segments. Our evaluation task has been limited to only the second-pass rescoring due to the computational complexity in search with use of context-dependent feature units. In addition to reporting the comparative evaluation results, we also provide a formal description of recently improved key components of the recognizer. In particular, we publish for the first time a comprehensive set of feature spreading rules over all five articulatory feature dimensions.

2. RECOGNIZER COMPONENT I: THE PHONOLOGICAL/ARTICULATORY SPACE

Define a phonological space spanning over a total of M dimensions. Following the theory of articulatory phonology [1], each dimension in the phonological space can be made associated with one distinct articulatory structure (which we call articulatory feature). We assume that the d^{th} dimension, Θ_d, is characterized by N_d distinct values: $\Theta_d \in \{s_d^1, s_d^2, \cdots, s_d^{N_d}\}$, each indexed by a segmental linguistic unit of speech. One may think of the d^{th} articulatory feature as being in one of N_d states. It is reasonable to assume that the M individual features whose changes constitute the state evolution process within the phonological/articulatory space are independent of each other during speech production. A first order Markov chain $\Lambda_d = \{\pi_i^d, a_{ij}^d\}$ is employed to represent the state evolution process in the d^{th} dimension, where π_i^d and a_{ij}^d are initial state occupation probabilities and state transition probabilities of Λ_d, respectively.

Since articulatory state sequences are hidden and the acoustic observation is an integrated result of state occupations in all articulatory dimensions, a composite Markov chain $\Lambda = \{\pi_i, a_{ij}\}$ with state space $\Theta = \Theta_1 \times \Theta_2 \times \cdots \times \Theta_M$ is needed in order to characterise the relationship between the phonological/articulatory process and the acoustic observation. States in Λ are defined as $s_i = (s_1^{\mu_1^i}, s_2^{\mu_2^i}, \cdots, s_M^{\mu_M^i})$, where $\mu_d^i \in \{1, 2, \cdots, N_d\}$ is the index to one of the N_d distinct values associated with the d^{th} articulatory feature. Under the independence assumption, we have

$$\pi_i = \prod_{d=1}^{M} \pi_{\mu_d^i}^d \quad ; \quad a_{ij} = \prod_{d=1}^{M} a_{\mu_d^i \mu_d^j}^d. \tag{1}$$

In our current implementation of the articulatory feature based speech recognizer, five articulatory features (M=5) are employed: three for major articulators (lip, tongue blade and tongue dorsum) and two for secondary articulators (velum and larynx). The total number of dis-

0-7803-2431-5/95 $4.00

tinct values associated with each of the articulatory features is 5, 7, 20, 2 and 3, respectively. For the major articulatory features, context sensitive feature values are used. The total number of all possible states in A is thus $(5 \cdot 3) \cdot (7 \cdot 3) \cdot (20 \cdot 3) \cdot 2 \cdot 3 = 113,400$. (Factor 3 is due to context sensitivity.) Fortunately, not all of these composite articulatory states are reachable during speech production. With the constraints of speech production and knowledge of coarticulation, only a small subset of these states are needed for characterising acoustic realisation of the phonological/articulatory process. In our current recognizer, an articulatory state transition graph is dynamically constructed for an arbitrary phonemic transcription of a speech utterance. The total number of composite states in our recognizer is 3,822.

3. RECOGNIZER COMPONENT II: PHONEMIC TO ARTICULATORY-STATE MAPPING

Let the phonemic transcription be $f_1, f_2, \cdots, f_T$ for a speech utterance. The transcription is given in the recognizer's training phase and is a hypothesis in the second-pass testing phase given the N-best evaluation scenario. The process of mapping phonemic representation to the articulatory states involves the following steps:

(1) Map each phoneme f_i to a context-independent composite articulatory state s_i consisting of quintuple feature values. A mapping table for all the TIMIT labels (quasi-phonemes) can be found in [6]. Feature underspecification is incorporated in this mapping.

(2) Create a triplet (s_i^l, s_i^c, s_i^r) for each articulatory state s_i to characterise the coarticulatory effects from both left (look-ahead) and right (articulators' inertia) directions. s_i^c is the *center* state (context dependent if f_i contains underspecified feature(s)), and s_i^l, s_i^r take account of the *left* and *right* contextual factors, respectively, for f_i.

(a) The feature value in an articulatory feature dimension of s_i^l is set to be the first specified one found by searching to left $s_{i-1}, s_{i-2}, \cdots, s_{i-R_l}$ in that dimension. $R_l = 3$ is the search range implemented in our current recognizer.

(b) s_i^r is constructed in the same manner as s_i^l except that context searching is to the right and that the range of the search is set to $R_r = 4$. (Look-ahead coarticulation is stronger than carry-over one in English.)

(c) The feature value in each dimension of s_i^c is set to that of s_i if that feature of s_i is specified; s_i^c inherits a left or right adjacent phoneme's *specified* feature value if the corresponding feature of s_i is underspecified.

(3) Create articulatory state-transition graph G_i for f_i in which a set of composite articulatory states are made to realize transitions from s_i^l to s_i^c and from s_i^c to s_i^r.

(a) Apply articulatory feature spreading rules which dictate what kinds of feature spreading need be prohibited. A comprehensive set of such rules are summaried in Figure 1, where each sub-figure covers one manner class of English sounds. A horizontal box which intrudes into the middle column from left/right column represents a rule that feature spreading from the left/right context in that feature dimension is possible. When feature spreading is prohibited according to the rules, the corresponding feature value of s_i^l or s_i^r is simply replaced by that of s_i^c.

(b) Construct *depth-zero* state(s) in G_i. If s_i^l differs from s_i^c by only one (or fewer) in the feature value among the three major articulatory feature dimensions, s_i^l will be the only depth-zero state of G_I. Otherwise, G_i will contain a set of depth-zero states, each of which inherits one major and both secondary articulatory feature values from s_i^l, and inherits the remaining two major articulatory feature values from s_i^c. Define state depth p=0 for the depth-zero state(s).

(c) Increment p by one. For all states in depth $p-1$, create all possible distinct states via replacing their feature values by the corresponding ones in s_i^c. Each of such replacements in each dimension will create one new state. Set the depth of all newly created states to p.

(d) repeat step (c) until only s_i^c is created.

(e) Increment p by one. For all states in depth $p-1$, create all possible distinct states via replacing their feature values by the corresponding ones in s_i^r. Remove any state that differs from s_i^c in more than one feature values among the three major articulatory dimensions. Set the depth of all newly created states as p.

(f) repeat step (e) until no new state can be created or until s_i^r is reached.

(g) A link (state transition) is created from any state of depth $k-1$ to that of depth k if these two states differ in only one of the five feature dimensions. Self loop is also created for each state.

(h) If f_i is a vowel, additional connections are created that skip s_i^c.

(4) Construct the composite articulatory state transition graph for $f_1, f_2, \cdots, f_T$ by combining $G_1, G_2, \cdots, G_T$. If f_i and f_{i+1} are both vowels or both consonants, G_i and G_{i+1} are combined by linking s_i^c to the depth-zero state(s) of G_{i+1} and by linking the maximum-depth state(s) of G_i to s_{i+1}^c. Otherwise, these two graphs are concatenated directly by linking the maximum-depth state(s) of G_i to the depth-zero state(s) of G_{i+1}.

4. RECOGNIZER COMPONENT III: ARTICULATORY-STATE TO ACOUSTIC MAPPING

For each composite articulatory state generated according to the above steps from phonemic transcriptions in the training corpus, a statistical distribution is used to cover variabilities in the acoustic observations conditioned on the state. On the other hand, the well-known phenomena of many-to-one mapping from articulation to acoustics can be accommodated via appropriately tying articulatory states so that these states (representing distinct articulatory configurations) nevertheless share the same distributional parameters that characterise speech acoustics.

In our current feature-based speech recognizer, all the Markov states have clear physical interpretation in terms of the underlying articulatory structures responsible for generating the acoustic observations. It is thus possible to use different parametric forms in statistical distributions to implement the acoustic mapping component of the recognizer. We have currently implemented both stationary and nonstationary statistical distributions for the articula-

tory states marked by assimilated features in one or more major articulatory feature dimension(s). Due to the computational difficulties encountered by the nonstationary version of the system, the experimental results presented in this paper are only from the stationary version, where the state-conditioned acoustic output distribution is i.i.d. mixture Gaussian densities.

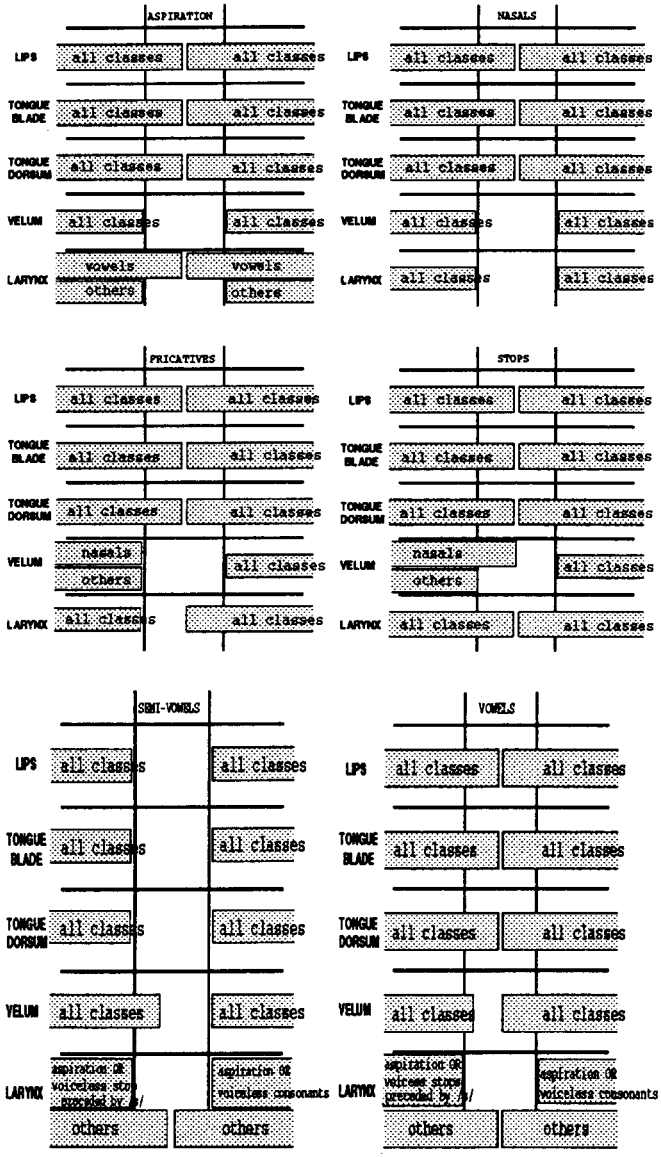

Figure 1. Articulatory Feature Spreading Rules

5. EXPERIMENTAL RESULTS

The feature-based speech recognizer described above has been evaluated in an experiment for phonetic recognition of standard 39 folded phone classes in continuous TIMIT sentences. An N-best candidate list is provided for each test utterance and the task is to re-order the candidate list according to the likelihoods computed from our feature-based speech recognizer. These N-best lists, as described in detail in [7], were generated in Cambridge University using HTK [9]. The lists contain up to 30 candidate sentences (given as phonetic strings) for each test utterance.

The model parameters in our feature-based speech recognizer are trained by performing 10 iterations of the segmentational K-means training algorithm on 3,696 sentences from 462 TIMIT recommended training speakers (SA sentences are excluded in training). 3,822 articulatory states are created to form state transition graphs for all training sentences. However, only 2,761 of them are made associated with acoustic observations; each of the remaining 1,061 states appears less than 4 times in the training corpus. The acoustic observation of each state is modeled by a mixture of 5 Gaussian densities. The covariance matrices of all Gaussian densities with fewer than 15 training frames are tied.

The state transition graph for a testing utterance may contain *unseen* articulatory states which have not appeared in the training corpus or have no acoustic observation data associated with them. If an unseen state has another states being in parallel with it, we simply remove it from the state transition graph. Otherwise, a null transition is created to skip this state.

The performance of the feature-based system is compared with that of a system constructed using generalized triphone models. Each generalized triphone is modeled by a three-state, left-to-right HMM with no state skipping. The middle state of the HMM is dependent only on the center phoneme of the triphone and the left/right state is dependent only on the left/right context of the triphone. In addition, center phonemes of the triphones are tied into 39 classes, and left/right context phonemes are tied into 15 classes. In this way, only 1,209 states, each with a mixture of 5 Gaussian densities, are used in the system. The 15 classes used for the merged contexts are obtained by modifying similar classes published in [4] and in [8]. The parameters of the generalized triphone models are estimated using the same training data and the same training algorithm as those for the feature-based models.

The testing set consists of 48 randomly selected SX sentences from 48 speakers (the selection process guarantees that each region has four male speakers and two female speakers). Table 1 and Table 2 gives detailed recognition performances for the feature-based and the generalized-triphone-based systems. Table 3 gives the performances of the worst and the best candidate sentences in the N-best list (the original candidate order in the N-best list can not be used to compare our results since that list was obtained using a word lexicon and word-pair grammars). Figure 2 plots the correct recognition rates and recognition accuracies as a function of the top number of candidate sentences among the 30 N-best ones for the two systems, respectively.

The above experimental results show that although the improvement in performance on the top-one candidate sentence of the feature-based system over that of the triphone-based system is marginal, the overall quality of the rescored N-best list from the feature-based system is clearly better than that from the triphone-based system. Greater performance improvements are observed as the top number of candidate sentences increases.

No. Cand.	Corr.	Acc.	Sub.	Del.	Ins.
1	78.62%	70.69%	6.29%	5.08%	7.93%
5	81.72%	74.47%	3.38%	4.89%	7.25%
10	83.40%	75.84%	1.90%	4.71%	7.56%
15	83.58%	76.21%	1.40%	4.96%	7.43%

Table 1. Performance of the feature-based based system

No. Cand.	Corr.	Acc.	Sub.	Del.	Ins.
1	77.57%	70.82%	7.22%	5.20%	6.75%
5	80.55%	73.79%	4.56%	4.89%	6.75%
10	81.47%	74.35%	3.75%	4.77%	7.13%
15	82.47%	75.40%	2.70%	4.83%	7.06%

Table 2. Performance of the triphone-based system

Choice	Corr.	Acc.	Sub.	Del.	Ins.
worst	67.16%	50.99%	26.70%	6.13%	16.17%
best	83.58%	76.39%	11.40%	5.02%	7.19%

Table 3. Worst and best performance in the N-best list

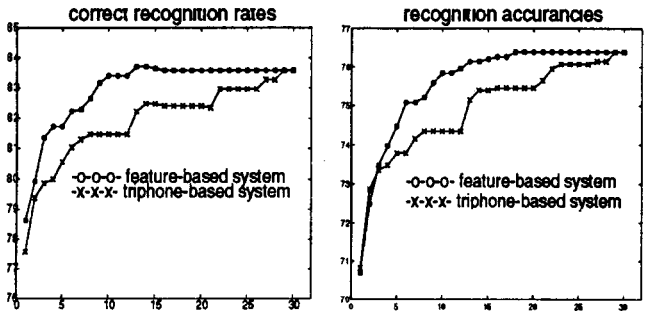

Figure 2. System performances as a function of the top number of candidate sentences.

6. SUMMARY AND DISCUSSIONS

In this paper, three key components of the articulatory feature based speech recognizer, which has been under development in our research laboratory in recent years, are described formally: the phonological/articulatory state space, the phonemic (discrete) to articulatory-state (discrete) mapping, and the articulatory-state (discrete) to acoustic-stream (continuous) mapping. The above first component stays essentially the same as that presented in [6]. The above second component, as a main focus of this paper, has been improved significantly recently over that published in [6]. The above third component has also been improved significantly over that published in [6] and has been described in a much greater detail in [3, 2]. Moreover, the evaluation results on the TIMIT phonetic recognition task presented in this paper add further evidence for the effectiveness of the articulatory-feature based approach demonstrated earlier in [6] from a comparatively simpler phonetic classification task.

One conspicuous trait of our current feature-based approach to speech recognition as described in this and our earlier publications is its unified treatment of phonological and articulatory (symbolized) representations of speech following the theory of articulatory phonology (for an overview article of the theory, see [1]). This perhaps simplifies certain phonological reality that would demand clear separation between the relatively higher level phonological process and articulation at the lower phonetic level. Another distinguishing quality of our approach is its rather unelaborate and naive treatment of the relationship between the articulatory representation of speech and its acoustic counterpart. This much more complicated relationship in reality has been functionally modeled in our current approach by a simple mapping from the discrete-state Markov chain (i.e. articulatory-state sequence) to the continuous-valued output observation in the HMM. One inevitable consequence resulting from the simplicity of our approach on both of the above accounts is the need to form Cartesian product, at least theoretically, in designing the the articulatory-state space. At present, we use coarticulation rules to control the size of the state space, enabling successful construction of our current functional speech recognizer. It appears that the adoption of the highly simplified treatments of both the phonology-to-articulation and articulation-to-acoustics relationships, as exemplified in construction of the speech recognizer described in this paper, is an effective way to enable us to design a functional, high-performance speech recognizer operative on all classes of speech sounds that would be otherwise impossible.

ACKNOWLEDGEMENTS

We are grateful to Matthew Jones of the Cambridge University, UK, for providing the N-best candidate lists of the testing utterances.

REFERENCES

[1] C. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, Vol.49, pp. 115-180, 1992.

[2] L. Deng and H. Sameti. "Transitional speech units and their representation by the regressive Markov states structured via assimilation of major articulatory features," *IEEE Trans. on Speech and Audio Proc.*, 1994, submitted.

[3] L. Deng and H. Sameti. "Speech recognition using dynamically defined speech units," *Proceedings of the 1994 International Conference on Spoken Language Processing*, Vol. 4, pp. 2167–2170, Yokohama, Japan, September, 18-22, 1994.

[4] L. Deng, M. Lennig, F. Seitz and P. Mermelstein. "Large vocabulary word recognition using context-dependent allophonic hidden Markov models," *Computer Speech and Language*, Vol. 4, 1990, pp. 345-357.

[5] L. Deng and D. Sun. "Phonetic classification and recognition using HMM representation of overlapping articulatory features for all classes of English sounds," *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, Adelaide, Australia, 1994, pp. 45-48.

[6] L. Deng and D. Sun. "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *J. Acous. Soc. Am.*, Vol. 95, No. 5, May 1994, pp. 2702–2719.

[7] M. Jones and P.C. Woodland, "Modeling syllable characteristics to improve a large vocabulary continuous speech recognizer," *Proc. ICSLP 94*, Yokohama, Japan, 1994

[8] A. Ljolje, "High accuracy phone recognition using context clustering & quasi-triphone models," *Computer, Speech & Language*, Vol.8, pp.129-151, 1994

[9] S.J. Young, "The HTK Hidden Markov Model toolkit: design and philosophy," Cambridge University Engineering Department Technical Report CUED/F-INFENG/TR.152., 1993.