Speech Attribute Recognition using Context-Dependent Modeling

Van Hai Do*†, Xiong Xiao†, Ville Hautamäki‡§, Eng Siong Chng*†
* School of Computer Engineering, Nanyang Technological University, Singapore
† Temasek Laboratories@NTU, Nanyang Technological University, Singapore
‡ Institute for Infocomm Research, Singapore
§ School of Computing, University of Eastern Finland, Finland
{dova0001, xiaoxiong, aseschng}@ntu.edu.sg, villeh@cs.joensuu.fi

***Abstract*—Speech attributes, such as places and manners of articulation are robust against cross-speaker variation and environmental distortions. They have been used in various speech processing applications such as spoken language identification, speaker recognition and speech recognition. In this paper, we propose a method to recognize speech attributes by using a context-dependent modeling of the attributes, called bi-attributes. Experimental results on the TIMIT database show that the context-dependent modeling reduces frame classification error by 13.2% and 16.1% relatively over the context-independent modeling for manner and place classification, respectively. In addition, when fused with phone posteriors to improve phone recognition accuracy, the attribute context dependent modeling gives a 9.9% relative phone error rate reduction over the attribute context independent modeling.**

I. Introduction

Speech attributes such as voicing, nasal, dental, describe how speech is articulated. They are also called phonological features, articulatory features, or linguistic features. It is known that speech attributes are robust against inter-speaker variation and not sensitive to environmental noise [1], [2]. In addition, speech attributes have been shown to be complementary to traditional features such as MFCCs or PLPs in automatic speech recognition [1], [2], [3].

Speech attributes have been used in various speech processing applications. In [4], attributes were used for spoken language identification. There, multiple parallel sequences of attributes were scored with separate feature n-gram models. In [5], attribute streams were used to construct conditional pronunciation models for speaker recognition. In the *automatic speech attribute transcription system* (ASAT) [6], a bank of attribute detectors were used as the front-end module. In [3], [7], manners and places of articulation were used as speech attributes to improve the accuracy of the phone recognition systems. In [8], posterior probabilities were estimated for 64 attributes and then used as the feature for an HMM/GMM system. Speech attributes were also used to rescore the lattice given by a large-vocabulary continuous speech recognition (LVCSR) system [9]. In [2], speech attributes were used to supplement HMMs of LVCSR tasks by combining the likelihood probability at phone and state levels. A comprehensive review of using speech attributes in speech recognition can be found in [10].

Approaches to attribute detection can be divided into two categories: frame based detection and segment based detection [11]. Segment based detectors can be implemented by HMMs, they are more reliable in spotting segments of speech. However, the detection curves are not synchronized in time [11]. In contrast to segment based detectors, frame based detectors can be realized with artificial neural networks (ANNs) [1], [3], [9], [12] or support vector machines (SVMs) [13]. The advantage of the frame based approach is that the outputs of the detector can approximate the speech attribute posterior probabilities given the speech signal [11]. In this paper, we examine frame based detectors implemented by ANNs as they are both flexible and effective. Specifically, a group of multilayer perceptron neural networks (MLPs) is used to classify the entire attribute subset.

In the reported attribute detectors [1], [3], [9], [12], [13], speech attribute posteriors were computed without any context information. The number of classes is the same as that of the attributes, which is typically small. In this paper, we propose detailed attribute classes by taking into account the context information of attributes. Specifically, attributes are expanded to bi-attributes by considering the left and the right context of the attributes. The number of context-dependent bi-attributes is therefore increased over that of context-independent mono-attributes. With the increased number of classes, feature variation in each class will be reduced. A similar approach has been used to model context-dependent phonemes (e.g., triphone) and obtained significant improvements in speech recognition [14].

The rest of this paper is organized as follows. Section II describes how context dependence can be applied in attribute recognition. The experimental results are reported in Section III, and Section IV concludes the paper.

II. Context Dependent Attribute Recognition

A. Definition of bi-attributes

In this paper, we focus on classifying manners and places of articulation [3] which include:

- Manners: vowel, stop, fricative, approximant, nasal, silence.
- Places: low, mid, high, dental, labial, coronal, retroflex, velar, glottal, silence.

	Sil	HELLO				Sil
Phones	**Sil**	**hh**	**ah**	**l**	**ow**	**Sil**
Manners	Sil	**F**ricative	**V**owel	**A**pproximant	**V**owel	Sil
Left-bi-manners	Sil	Sil-**F**	F-**V**	V-**A**	A-**V**	Sil
Right-bi-manners	Sil	**F**+V	**V**+A	**A**+V	**V**+Sil	Sil
Places	Sil	**G**lottal	**M**id	**C**oronal	**M**id	Sil
Left-bi-places	Sil	Sil-**G**	G-**M**	M-**C**	C-**M**	Sil
Right-bi-places	Sil	**G**+M	**M**+C	**C**+M	**M**+Sil	Sil

Fig. 1. *Assignment of bi-manner and bi-place labels to a sequence of speech segments.*

To train the attribute and bi-attribute detectors with MLPs, each speech frame in the training data needs a label, i.e., the ground truth of which manner and place classes the frame belongs to. Figure 1 illustrates how attribute labels can be created from phone labels. From phonetics, we know that each phone is a combination of one manner and one place [1]. Hence manner and place attributes can be simply generated from the phone label as shown in the figure. For example, the manner and place of phone "hh" are "Fricative" and "Glottal", respectively.

The labels for bi-manners and bi-places are generated from the labels of manners and places respectively by taking into account the labels of neighboring attributes. We use two types of bi-attributes, the left context bi-attributes and the right context bi-attributes. The reason for doing so is that an attribute is usually affected by its previous attribute and following attribute. Ideally, we could use a tri-attribute label, similar to what is done in triphone modeling in speech recognition [14]. However, the number of tri-attributes will be too big and the computation requirement will be too heavy. Hence, we use two types of bi-attributes to model the left context and the right context with a modest increase of the total number of bi-attribute classes.

The generation of bi-attribute labels is also illustrated in Figure 1. For example, to generate the left-bi-manner label of phone "ah", the manner label of its previous phone "hh" is also used, which results in "F-V" representing "**F**ricative-**V**owel". The label "F-V" is assigned to all the frames belong to phone "ah" when we train the left-bi-manner detector. Similarly, the right-bi-manner label of "ah" will be "V+A" which represents "**V**owel" followed by "**A**pproximant". The generation of the left and the right bi-place labels is similar. Silence is kept as an independent attribute without context.

Four detectors are used for bi-attribute modeling, they are left-bi-manner, right-bi-manner, left-bi-place and right-bi-place detectors. Given a speech feature vector, a detector does not only give the single best class, but also posterior probabilities of all classes.

B. Attribute recognition

We use manner and place detectors to produce the baseline results for attribute recognition. In another way, attribute posterior probabilities are estimated from bi-attribute detectors by marginalizing all bi-attribute posterior probabilities which correspond to the examining attribute. Equation (1) illustrates how attribute posterior probabilities are computed from the left-bi-attribute posterior probabilities. A summation is used since bi-attributes are not overlapped classes.

$$P^{\text{left}}([A_i]|o(t)) = \sum_{j=1}^{N_i^{\text{left}}} P([A_j - A_i]|o(t)). \quad (1)$$

where: $o(t)$ is the feature vector; $[A_i]$ is the considered attribute; $[A_j - A_i]$ is the left-bi-attribute of the current attribute $[A_i]$ and the left context attribute $[A_j]$; N_i^{left} is the number of the left-bi-attributes for $[A_i]$.

Similarly, attribute posterior probabilities are also estimated from the right-bi-attribute posterior probabilities as:

$$P^{\text{right}}([A_i]|o(t)) = \sum_{j=1}^{N_i^{\text{right}}} P([A_i + A_j]|o(t)). \quad (2)$$

We combine both contextual attribute posteriors by simply taking the average value (unweighted sum rule [1]) as:

$$P([A_i]|o(t)) = \frac{P^{\text{left}}([A_i]|o(t)) + P^{\text{right}}([A_i]|o(t))}{2}. \quad (3)$$

III. Experiments

A. Experimental setup

Database: The TIMIT database[1] is used in our experiments. The SA part of the TIMIT database is not used. The training set consists of 3696 utterances from 462 speakers. A small part extracted from the training set (50 speakers) is used as the development set. The complete test data set contains 1344 utterances from 168 speakers.

Phone set and bi-attribute set: The original 64 phonetic labels are mapped into the 45 phones as described in [15]. Phones "cl", "vcl", and "epi" are merged into the phone "sil". We use the mapping from 45 phones into 6 manners and 10 places [16]. There are 31 left-bi-manners, 31 right-bi-manners, 89 left-bi-places, and 89 right-bi-places. For phone recognition results, after decoding, the phone accuracies are computed by down mapping the recognition output from 45 phones to 39 phones [15].

Detector architecture: There are seven detectors used: two detectors for attributes, four detectors for bi-attributes and one detector for generating phone posteriors. In our experiments, all detectors have the same structure which is built by three MLPs with one hidden layer as shown in Figure 2 [7]. A sliding window of 31 speech frames around the current frame is used and divided into two equal parts, 16 frames in the left side and 16 frames in the right side (the current frame is overlapped). In each part, 15 critical bands are used and each is represented by 11 DCT (Discrete Cosine Transform) coefficients. Hence, there is a total of 15x11=165 inputs for the left or right MLP.

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

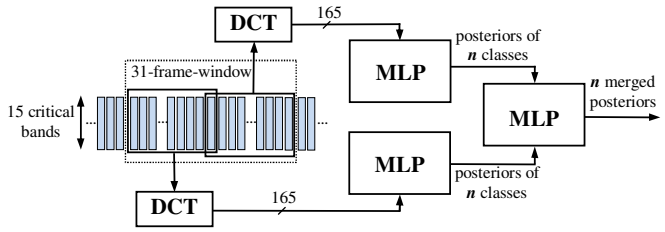


Fig. 2. *Block diagram of a detector, n is the number of classes, e.g., attributes, bi-attributes, or phones, depending on the detector is an attribute or bi-attribute or phone detector.*

Software: The ICSI QuickNet software package[2] is used to employ MLPs with one hidden layer. The Viterbi algorithm used to produce the recognized phone strings is implemented by HVite, and HResults in the HTK speech toolkit[3] is used to evaluate the phone recognition results.

Parameter selection criteria: The training process of MLPs is stopped to avoid over-fitting when the classification error of the development set starts to increase. The number of hidden unit in each MLP is examined in a range from 200 to 3000 and the MLP which produces the lowest error rate on the development set is selected. In the decoding process, the phone insertion penalty value is also tuned based on the development set.

B. Attribute recognition

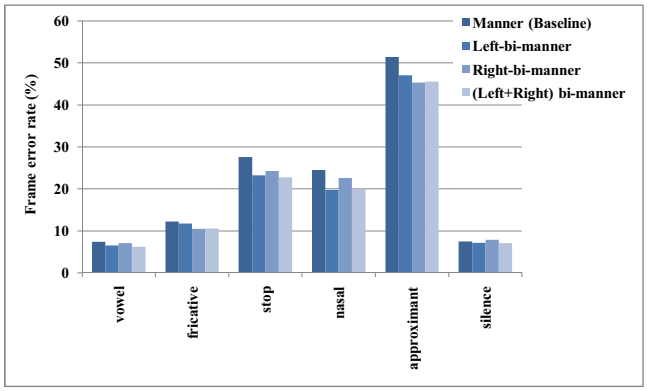
In this study, attribute recognition performance is measured by *frame error rate* (FER) of classification. For each frame, the attribute with the highest posterior probability is recognized as the attribute for the frame. Table 1 shows FERs of attribute recognition for manners and places. For comparison, we also list the results reported in [3], [12]. Note that our baseline system produces different results from [3], [12]. This may be due to three reasons: 1) we use features of long temporal contexts (310ms) instead of short time features (e.g., MFCCs), 2) the number of hidden units in each MLP is optimized in our experiments, and 3) bi-gram feature models and durational models used in [12] to improve recognition results are not used in this paper.

From the table, it is observed that the context-dependent modeling (the last 3 rows) outperforms the context-independent modeling ("Attribute baseline") significantly for both manner and place recognition. In addition, the fusion of the left and the right context manner and place posteriors improve the performance further, which shows that the left context and the right context provide complementary information. The combined systems (Equation 3) in the last row reduces the FERs by 13.2% and 16.1% over the context-independent baseline for manner and place recognition, respectively. Further improvement can be obtained if more complicated combination schemes are used (e.g., weighted sum, nonlinear combinations).

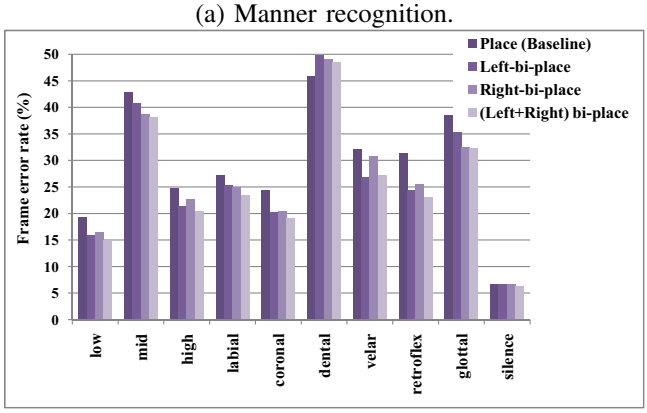
[2]ICSI QuickNet package, http://www.icsi.berkeley.edu/Speech/qn.html
[3]HTK speech toolkit, http://htk.eng.cam.ac.uk

TABLE I
FRAME ERROR RATES (FERS) OF ATTRIBUTE RECOGNITION. THE NUMBERS AFTER "/" IN THE LAST 3 ROWS ARE THE RELATIVE FER REDUCTION ACHIEVED OVER THE "ATTRIBUTE BASELINE".

Method	**Manner** (in %)	**Place** (in %)
Context-independent Modeling		
Results from [3]	17.90	26.80
Results from [12]	13.50	27.00
Attribute baseline	14.05	22.74
Context-dependent Modeling		
Left-bi-attribute	12.65/10.0	20.16/11.4
Right-bi-attribute	12.95/7.8	20.33/10.6
Left+Right bi-attribute	12.19/**13.2**	19.07/**16.1**



(a) Manner recognition.



(b) Place recognition.

Fig. 3. Comparison of attribute recognition performance using the attribute detectors and the bi-attribute detectors.

FERs for each attribute are plotted in Figure 3. From the figure, it is observed that FERs of the left-bi-attribute detectors and the right-bi-attribute detectors are lower than FERs of the attribute detectors in all cases except "dental" and "silence". The results also show that the combination of the left and the right bi-attribute posteriors produces consistently better performance (except "nasal", "approximant" and "velar").

Figure 4 gives an example of manner posteriors estimated by manner and bi-manner detectors. The top panel shows the ground truth, the middle panel shows the posteriors generated by the context-independent manner detector, and the bottom panel shows the posteriors obtained by the fusion of the left and the right bi-manner detectors. We can observe from the figure that the posteriors generated by the bi-manner detectors

are clearly closer to the ground truth than the manner detector.

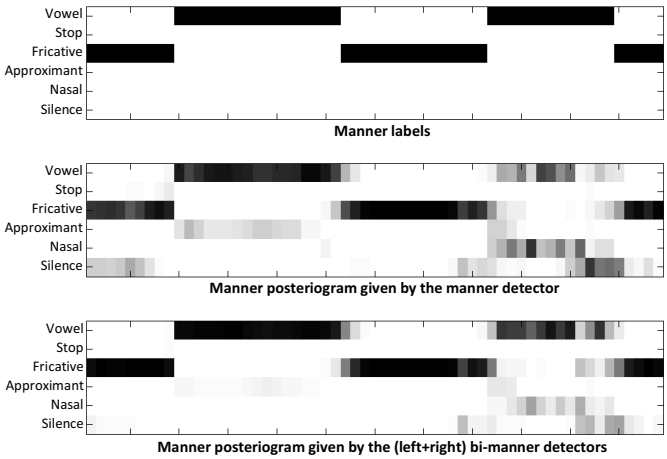


Fig. 4. Comparison of manner posteriors estimated by the manner detector and the (left+right) bi-manner detectors.

C. Phone recognition

In this section, we carry out simple experiments to demonstrate the usefulness of attributes and bi-attributes in phone recognition. Five experiments have been conducted as shown in Figure 5 and the experimental results are listed in Table 2. Note that all systems use the same original feature.

After phone posterior probabilities are computed, they are converted into scaled likelihoods by using the Bayes formula and we assume that the prior probabilities of all phones are equal. We then use HVite to decode them into phone sequences. The phone insertion penalty is used to control the decoding process. This approach has been applied successfully in the hybrid HMM/ANN systems [17].

TABLE II
PHONE ERROR RATES (PERS) FOR PHONE RECOGNITION (NO LANGUAGE MODEL, 1-STATE-MONOPHONE MODELS).

Experiment	Method	PER (in %)
(1)	Phones	30.29
(2)	Phones + attributes (baseline)	29.83
(3)	Phones + left-bi-attributes	27.99
(4)	Phones + right-bi-attributes	28.15
(5)	Phones + (left+right) bi-attributes	26.89

In experiment (1), phone posteriors are estimated directly from the phone detector (which has the same architecture as attribute and bi-attribute detectors). In experiment (2), initial phone posteriors from experiment (1) are concatenated with manner and place posteriors given by the context-independent manner and place detectors and mapped to phone posteriors by an MLP. The result of this experiment shows that manners and places of articulation can be used to improve speech recognition performance. It is also consistent with previous research in [1], [2], [3], [7], [9], [10], [16]. Experiment (3) and (4) are the cases where initial phone posteriors are fused with the left and the right bi-attribute posteriors, respectively.

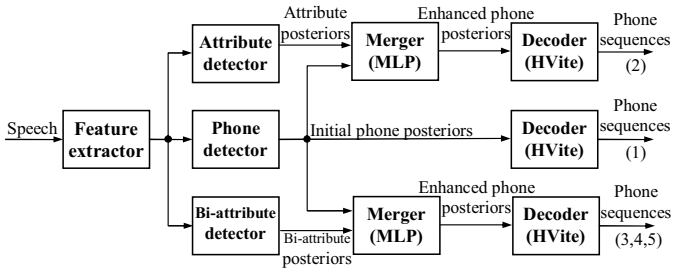


Fig. 5. *Block diagram of different phone recognition systems.*

Significant improvements are achieved in (3) and (4) over (2). Furthermore, in experiment (5) where initial phone posteriors are fused with both left and right bi-attribute posteriors, further improvement is observed over (3) and (4). The best result in (5) represents a 9.9% relative PER reduction over the baseline in (2). These results show that bi-attributes provide significantly more complementary information than mono-attributes for phone recognition task.

Also note that none of our phone recognition results approaches the best result on TIMIT, 20.96% PER [7]. Our experiments are built to illustrate the difference between attributes and bi-attributes in improving phone recognition. In addition, in our experiments, no language model is used and all phones are modeled as single-state-monophones.

IV. CONCLUSION

In this paper, we proposed a context-dependent modeling for better speech attribute recognition. Results on the TIMIT database confirmed that context information is helpful for improving attribute recognition. In addition, we also show that using context information in bi-attribute modeling produces significantly more complementary information than without using context information for phone recognition task.

We believe that the better attribute recognition will also benefit other speech processing applications. For the future work, we will examine ways to increase the context size by considering both the left and the right context of attributes jointly as well as to apply our method for different knowledge sources.

V. ACKNOWLEDGEMENT

The authors would like to thank Dr. Dau-Cheng Lyu of Temasek Laboratories@NTU, Nanyang Technological University for help on setting up the baseline attribute detector system used in this paper.

REFERENCES

[1] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *ICSLP*, 1998, pp. 891-894.

[2] F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features," in *ICSLP*, 2002.

[3] J. Li, Y. Tsao, and C.-H. Lee, "A study on knowledge source integration for candidate rescoring in automatic speech recognition," in *ICASSP*, vol. I, 2005, pp. I837-I840.

[4] K. Kirchhoff and S. Parandekar, "Multi-stream statistical N-gram modeling with application to automatic language identification," in *EuroSpeech*, 2001, pp. 803-806.
[5] K. Leung, M. Mak, and S. Kung, "Articulatory feature-based conditional pronunciation modeling for speaker verification," in *ICSLP*, 2004, pp. 2597-2600.
[6] C.-H. Lee, M. A. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, B. Juang, and L. R. Rabiner, "An overview on automatic speech attribute transcription (ASAT)," in *InterSpeech*, 2007, pp. 513-516.
[7] S. Siniscalchi, P. Schwarz, and C.-H. Lee, "High-accuracy phone recognition by combining high-performance lattice generation and knowledge based rescoring," in *ICASSP*, vol. 4, 2007, pp. IV869-IV872.
[8] O. Cetin, A. Kantor, S. King, C. Bartels, M. Magimai-Doss, J. Frankel, and K. Livescu, "An articulatory feature-based tandem approach and factored observation modeling," in *ICASSP*, vol. 4, 2007, pp. 645-648.
[9] S. Siniscalchi, T. Svendsen, and C.-H. Lee, "A phonetic feature based lattice rescoring approach to LVCSR," in *ICASSP*, 2009, pp. 3865-3868.
[10] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, pp. 723-742, 2007.
[11] J. Li and C.-H. Lee, "On designing and evaluating speech event detectors," in *InterSpeech*, 2005, pp. 3365-3368.
[12] K. Hacioglu, B. Pellom, and W. Ward, "Parsing speech into articulatory events," in *ICASSP*, 2004, pp. 925-928.
[13] U. Chaudhari and M. Picheny, "Articulatory feature detection with Support Vector Machines for integration into ASR and phone recognition," in *ASRU*, 2009, pp. 93-98.
[14] S. Young, "Large vocabulary continuous speech recognition: A review," *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 45-57, 1996.
[15] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641-1648, 1989.
[16] S. Siniscalchi and C.-H. Lee, "A study on integrating acousticphonetic information into lattice rescoring for automatic speech recognition," *Speech Communication*, vol. 51, no. 11, pp. 1139-1153, 2009.
[17] H. Bourlard and N. Morgan, "Continuous speech recognition by connectionist statistical methods," *IEEE Transactions on Neural Networks*, vol. 4, no. 6, pp. 893-909, Nov. 1993.