# Automatic Speech Recognition and Dependency Network to Identification of Articulation Error Patterns

Yeou-Jiunn Chen, Jiunn-Liang Wu, and Hui-Mei Yang

***Abstract*—Articulation errors will seriously reduce speech intelligibility and the ease of spoken communication. Typically, a language therapist uses his or her clinical experience to identify articulation error patterns, a time-consuming and expensive process. This paper presents a novel automatic approach to identifying articulation error patterns and providing pronunciation error information to assist the linguistic therapist. A photo naming task is used to capture examples of an individual's articulation patterns. The collected speech is automatically segmented and labeled by a speech recognizer. The recognizer's pronunciation confusion network is adapted to improve the accuracy of the speech recognizer. The modified dependency network and a multiattribute decision model are applied to identify articulation error patterns. Experimental results reveal the usefulness of the proposed method and system.**

## I. Introduction

ARTICULATION errors, which generate different degrees of abnormality in articulation, seriously reduce speech intelligibility and the ease of spoken communication. Typically, a speech-language pathologist uses his or her clinical experience to identify articulation error patterns, a time-consuming and expensive process. Therefore, an automatic process for identification of articulation error patterns is very helpful to assist speech-language pathologist in clinical speech evaluation.

Most articulation errors fall into those three categories: omissions, substitutions, or distortions. For speech-language pathologist, the articulation errors are examined in terms of the place and manner of articulation and can be classified into five articulation error patterns: fronting, backing, de-aspiration, stopping, affrication, and omission [1]. In a typical fronting error, for example, a child may say /t/ instead of /k/ in the Chinese word /kan4/ so it would be heard as /tan4/. For backing error, the /q/ will be pronounced as the /k/ and /qi4/ would be heard as /ki4/.

Recently, researches usually aim at computer assistant treatment by using visual feedback [2, 3], which is useful to improve articulation ability. However, those approaches focus on design of corpus for training and treating process. The identification of articulation error patterns and pronunciation error information cannot be provided to a speech-language pathologist.

Other researchers proposed various approaches to identify articulation errors using statistical models [4] or tongue detection models [5-7]. For statistical models, Georgoulas et al. applied support vector machine to classify only three consonant phonemes. Only those phonemes were insufficient to identify the articulation error patterns. For tongue detection models, using ultrasound to examine speech production is gaining popularity because of its portability and noninvasiveness. The place of tongue could be detected from the ultrasound image. However, the resolution of detected results was insufficient to distinguish the five articulation error patterns. Moreover, the manner of articulation is also cannot be detected by this approaches.

Automatic speech recognition had been applied to many applications [8, 9] and the articulation attributes can be effectively estimated by speech technology [10, 11]. Therefore, it would be very useful to identify the pronunciation error information. Besides, dependency network (DN) technique was also applied to collective classification and suitable to knowledge discovery [12, 13]. From the clinical practice and experience, the information for identification of articulation error pattern errors is dependency and dependency network technique is appropriate to identify articulation error patterns.

In this paper, a novel automatic approach integrating automatic speech recognition and dependency network is proposed to identify articulation error patterns. A photo naming task (PNT) is used to capture examples of an individual's articulation patterns. The collected speech is automatically segmented and labeled by automatic speech recognition. Besides, the recognizer's pronunciation confusion network (PCN) is adapted to improve the accuracy of the speech recognizer. DN and multiattribute decision model are applied to estimate the likelihood of articulation error pattern by integrating the information of testing phoneme, labeled results, and speech.

## II. Methods

As shown in Fig. 1, the articulation disorder is actuated to articulate the names in photo naming task and the examples in speech are captured. For the speech, the automatic speech recognition is applied to segment and label it. Then, the labeling result and corresponding likelihood for each phoneme are used to identify articulation error patterns by DN and multiattribute decision model.

Manuscript received December 31, 2007. This work was supported in part by the National Science Council, Republic of China under Grant NSC95-2221-E-218-002-MY2.

Yeou-Jiunn Chen is with the Department of Electrical Engineering, Southern Taiwan University, Tainan County, Taiwan, R.O.C. (phone: 886-6-2533131 ext. 3325, fax: 886-6-3010073, e-mail: chenyj@mail.stut.edu.tw).

Jiunn-Liang Wu and Hui-Mei Yang are with Department of Otolaryngology, National Cheng Kung University Hospital, Tainan, Taiwan, R.O.C. (e-mail: jiunn@mail.ncku.edu.tw, tannas00@yahoo.com.tw).

## A. Photo Naming Task

To identify the articulation error pattern, speech-language pathologists design a PNT, which is composed of familiar vocabulary words that are represented by recognizable pictures. Let $w$ denote a recognizable picture and $s$ denote phoneme in $w$. Therefore, the collection of all phonemes in all $w$'s $S=\{s_1, s_2, \ldots, s_M\}$ and the total number of basic articulation units is $M$. For the words in PNT, three criterions should be satisfied. First, the word used in PNT should be a familiar vocabulary word with recognizable picture. This will decrease or eliminate the need for the child to imitate the clinician when presenting test stimulus items. Second, it should include the production of all phonemes. Those phonemes should be presented in at least two different word positions. Third, it should assess sounds in increasingly complex contexts. It should include target sounds in mono-syllabic and multi-syllabic words.

## B. Automatic Segmentation and Labeling

For each word $w_j = s_{j1}s_{j2}\cdots s_{jN_j}$, the $s_{jm}$ is the *m-th* phoneme and modeled by Hidden Markov Model (HMM) []. The corresponding speech, $o_j$ is used to automatically segment and label by maximum posterior probability as follows:

$$
\begin{aligned}
\hat{w}_j &= \hat{s}_{j1}\hat{s}_{j2}\cdots\hat{s}_{jN_j} \\
&= \underset{\tilde{s}_{j1}\tilde{s}_{j2}\cdots\tilde{s}_{jN_j}}{\arg\max}\ P\left(\tilde{s}_{j1}\tilde{s}_{j2}\cdots\tilde{s}_{jN_j}\left|o_j s_{j1} s_{j2}\cdots s_{jN_j}\right.\right) \\
&= \underset{\tilde{s}_{j1}\tilde{s}_{j2}\cdots\tilde{s}_{jN_j}}{\arg\max}\left(P\left(\tilde{s}_{j1}\tilde{s}_{j2}\cdots\tilde{s}_{jN_j}\left|s_{j1}s_{j2}\cdots s_{jN_j}\right.\right)^{\omega_l} P\left(\tilde{s}_{j1}\tilde{s}_{j2}\cdots\tilde{s}_{jN_j}\left|o_j\right.\right)^{\omega_a}\right)^{1/\omega_l+\omega_a} \\
&= \underset{\tilde{s}_{j1}\tilde{s}_{j2}\cdots\tilde{s}_{jN_j}}{\arg\max}\prod_m\left(P\left(\tilde{s}_{jm}\left|s_{jm}\right.\right)^{\omega_l} P\left(\tilde{s}_{jm}\left|o_{jm}\right.\right)^{\omega_a}\right)^{1/\omega_l+\omega_a} \\
&= \underset{\tilde{s}_{j1}\tilde{s}_{j2}\cdots\tilde{s}_{jN_j}}{\arg\max}\prod_m\left(P\left(\tilde{s}_{jm}\left|s_{jm}\right.\right)^{\omega_l}\left(P\left(o_{jm}\left|\tilde{s}_{jm}\right.\right)\frac{P\left(\tilde{s}_{jm}\right)}{P\left(o_{jm}\right)}\right)^{\omega_a}\right)^{1/\omega_l+\omega_a} \\
&= \underset{\tilde{s}_{j1}\tilde{s}_{j2}\cdots\tilde{s}_{jN_j}}{\arg\max}\prod_m\left(P\left(\tilde{s}_{jm}\left|s_{jm}\right.\right)^{\omega_l} P\left(o_{jm}\left|\tilde{s}_{jm}\right.\right)^{\omega_a}\right)^{1/\omega_l+\omega_a}
\end{aligned}
\tag{1}
$$

where $o_{jm}$ and $\tilde{s}_{jm}$ are the *m-th* speech segmentation and labeling results generated by Viterbi algorithm. $w_l$ and $w_a$ are the weighting factors for language and acoustic information. $P\left(o_{jm}\left|\tilde{s}_{jm}\right.\right)$ is the acoustic information estimated by HMM and $P\left(\tilde{s}_{jm}\left|s_{jm}\right.\right)$ is the language information estimated by maximum likelihood estimation (MLE) as

$$
P\left(\tilde{s}_m^j\left|s_m^j\right.\right) = \frac{P\left(\tilde{s}_m^j\right)}{P\left(s_m^j\right)} = \frac{C\left(\tilde{s}_m^j\right)}{C\left(s_m^j\right)} \tag{2}
$$

In order to improve the accurate of speech recognition, PCN as shown in Fig. 2 is used to guide the search space of Viterbi algorithm. The final state of each phoneme is connected to the collector state by a null transition, with probability 1. The collector state is then connected to the starting state by another null transition, with transition probability $P\left(\tilde{s}_{jm}^r \mid s_{jm}\right)$. $\tilde{s}_{jm}^r$ is the *r-th* phoneme candidate of $s_{jm}$.

## C. Identification of Articulation Error Patterns

For identification of articulation error patterns, a multiattributed decision model as shown in Fig. 3 is applied to integrate the likelihoods of all phonemes in PNT. For the *i-th* articulation error pattern, the likelihood of $E^i$ can be estimated multiattribute decision model and computed by the equation as follows:

$$
L\left(E^i\right) \propto \left(\frac{1}{M}\sum_m P\left(E_m^i s_m \hat{s}_m o_m\right)^\eta\right)^{\frac{1}{\eta}} \tag{3}
$$

where $\eta$ is a positive number and can be a coefficient to select competing decision. The final identified result can be decided by the posteriori probability when $L(E^i)>H_i$, where $H_i$ is a predefined threshold of $E^i$.

As shown in Fig. 4, a DN for a phoneme is established to model identification process of speech-language pathologists. However, the factor of articulation disorder is not considered for clinical practice and can be eliminated as shown in Fig. 5. The labeling process in PNT is also consulted by speech information $\hat{S}_m^a$ and language information $\hat{S}_m^l$ for speech-language pathologists. Thus, the DN for identification of articulation error pattern is shown in Fig. 6 and the corresponding likelihood can be estimated by the definition of DN as

$$
\begin{aligned}
P\left(E_m^i s_m \hat{s}_m o_m\right) &= P\left(E_m^i\left|s_m \hat{s}_m\right.\right) P\left(\hat{s}_m\left|s_m o_m\right.\right) P\left(o_m\right) P\left(s_m\right) \\
&\cong P\left(E_m^i\left|s_m \hat{s}_m\right.\right)\left(P\left(\hat{s}_m^l\left|s_m\right.\right)^{\omega_l} P\left(\hat{s}_m^a\left|o_m\right.\right)^{\omega_a}\right)^{1/\omega_l+\omega_a} P\left(o_m\right) P\left(s_m\right) \\
&\cong P\left(E_m^i\left|s_m \hat{s}_m\right.\right)\left(P\left(\hat{s}_m^l\left|s_m\right.\right)^{\omega_l}\left(P\left(o_m\left|\hat{s}_m^a\right.\right)\frac{P\left(\hat{s}_m^a\right)}{P\left(o_m\right)}\right)^{\omega_a}\right)^{1/\omega_l+\omega_a} P\left(o_m\right) P\left(s_m\right)
\end{aligned}
\tag{4}
$$

Since the $P(s_m)$ and $P(o_m)$ is the priori probability of a phoneme and a speech, it will slightly affect the likelihood for all articulation error pattern and can be eliminated to reduce the complicated. Thus, Eq. (4) can be derived as

$$
P\left(E_m^i s_m \hat{s}_m o_m\right) \cong P\left(E_m^i\left|s_m \hat{s}_m\right.\right)\left(P\left(\hat{s}_m^l\left|s_m\right.\right)^{\omega_l} P\left(o_m\left|\hat{s}_m^a\right.\right)^{\omega_a}\right)^{1/\omega_l+\omega_a} \tag{5}
$$

It is clearly that the probability $P\left(\hat{s}_m^l\left|s_m\right.\right)$ and $P\left(o_m\left|\hat{s}_m^a\right.\right)$ can be computed in the process of segmentation and labeling by automatic speech recognition. $P\left(E_m^i\left|s_m \hat{s}_m\right.\right)$ can be estimated

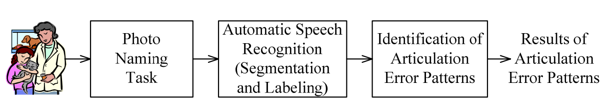


Fig. 1. The block diagram for identification of articulation error patterns.

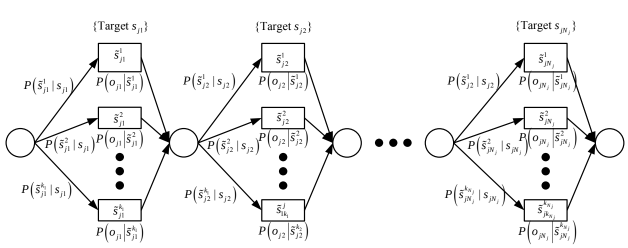


Fig. 2. The architecture of PCN for $w_j$

by MLE as

$$P\left(E_i \middle| s_m \hat{s}_m\right) = \frac{C\left(E_m\right)}{C\left(s_m \hat{s}_m\right)} \tag{6}$$

## III. Results and Discussion

Samples were collected from 553 children (346 males and 207 females) with multiple articulation error patterns. 421 and 132 samples were used for training and testing, respectively. The articulation error patterns of those samples were manually labeled by speech-language pathologists. In the training database, there are 45, 179, 88, 297, 106, and 42 samples for fronting, backing, de-aspiration, stopping, affrication, and omission, respectively. Moreover, in the testing database, there were 15, 57, 28, 95, 33, and 13 samples for fronting, backing, de-aspiration, stopping, affrication, and omission, respectively.

In Eq. (6), the priori probability of each pronunciation error is different to determine the articulation error patterns and should be estimated in the training database. The probability of distributions of pronunciation errors for articulation error patterns is shown in Table I. It is clear that the correlation between pronunciation error and articulation error pattern is different. Some pronunciation errors can give confident to identify a articulation error pattern. However, for clinical practice, to identify a articulation error pattern should be verify by different phoneme's pronunciation characteristic. Thus, the multiattribute decision model is proper to this mechanism.

To decide the identification results, a threshold of articulation error pattern should be determined. The receiver operating characteristic (ROC) curves for each identification results of articulation error patterns in training database were shown in Fig. 7. The equal error rates of Fronting, Backing, De-aspiration, Stopping, Affrication, and Omission were 7.32%, 11.78%, 9.87%, 8.76%, 7.07%, and 4.89%. Moreover, the thresholds with equal error rate were 0.16, 0.086, 0.092, 0.2, 0.2, and 0.1, respectively.

In order to evaluate the performance, the accuracy, specificity, sensitivity, and Kappa is used in this paper. The accuracy is number and proportion of all the observations in the table which have been classified correctly as

$$Accuracy = \frac{\left(TP + TN\right)}{\left(TP + TN + FP + FN\right)}. \tag{7}$$

The specificity of a test can be described as the proportion of true negatives it detects of all the negatives as

$$Specificity = \frac{TN}{\left(TN + FP\right)}. \tag{8}$$

The sensitivity of a test can be described as the proportion of true positives it detects of all the positives as

$$Sensitivity = \frac{TP}{\left(TP + FN\right)}. \tag{9}$$

The Kappa is a measure of agreement between predicted and observed as

$$Kappa = \left(P_o - P_e\right) / \left(1 - P_e\right) \tag{10}$$

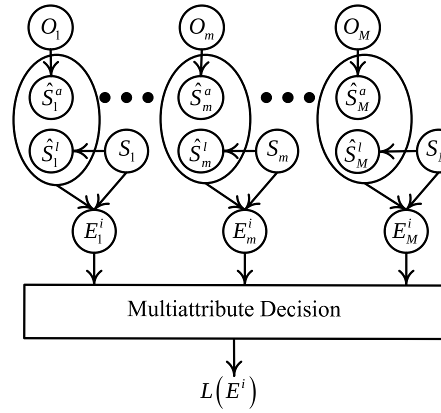


Fig. 3 The multiattribute decision model for identifying articulation error patterns.

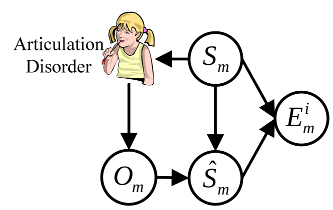


Fig. 4 DN for identification of articulation error pattern with articulation disorder.

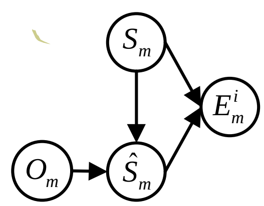

Fig. 5 DN for identification of articulation error pattern

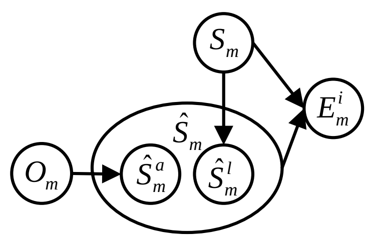

Fig. 6. DN for identification of articulation error pattern according to acoustic and language information

where

$$P_o = TP + TN \tag{11}$$

and

$$P_e = (TP+FN)(TP+FP) + (FN+TN)(FP+TN). \tag{12}$$

It takes on the value zero if there is no more agreement between test and outcome then can be expected on the basis of chance. Kappa takes on the value 1 if there is perfect agreement; i.e. the test always correctly predicts the outcome.

For the testing database, the accuracy, specificity, sensitivity, and Kappa for each articulation error pattern were shown in Fig. 7. Moreover, in order to discover the effect of the recognition error of automatic speech recognition. The labeling results in PAN were manually labeled by speech-language pathologists. The acoustic probability was also eliminated in DN and the results for identification of

TABLE I
PROBABILITY DISTRIBUTIONS OF PRONUNCIATION ERROR FOR EACH ARTICULATION ERROR PATTERN

| Fronting | | Backing | | De-aspiration | |
|---|---|---|---|---|---|
| PE | PB | PE | PB | PE | PB |
| g→d | 83.33 | d→g | 87.96 | p→b | 73.99 |
| k→d | 76.92 | t→k | 92.57 | t→d | 68.79 |
| k→t | 81.08 | zi→d | 16.94 | k→g | 7.41 |
| zhi→d | 18.80 | zi→g | 93.09 | q→j | 65.96 |
| chi→d | 27.48 | ci→d | 10.42 | ci→zi | 76.92 |
| chi→t | 15.85 | ci→t | 14.89 | | |

| Stopping | | Affrication | | Omission | |
|---|---|---|---|---|---|
| PE | PB | PE | PB | PE | PB |
| f→b | 75.26 | x→j | 66.15 | b→NULL | 87.50 |
| l→g | 90.91 | shi→zi | 81.58 | p→NULL | 45.61 |
| ri→g | 83.87 | si→zi | 70.97 | m→NULL | 38.89 |
| zi→d | 28.13 | | | f→NULL | 47.50 |
| zi→g | 96.28 | | | d→NULL | 71.79 |
| ci→d | 89.58 | | | t→NULL | 75.00 |
| ci→t | 87.23 | | | n→NULL | 23.26 |

PE: Pronunciation Error
PB: Probability

articulation error pattern were shown in Fig. 8. Although the accuracy of identification of articulation error pattern was degraded by the error labeling of automatic speech recognition, the system is used to assist speech-language pathologists to preliminarily examine articulatory ability. Thus, choosing a suitable threshold, it can achieve high specificity and low specificity and is useful to assist speech-language pathologists.

## IV. CONCLUSIONS

This work has presented an innovative approach to identify the articulation error patterns of articulation disorders. The examples of an individual's articulation pattern could be captured friendly by photo naming task. Automatic speech recognition is applied to automatic segment and label the articulation examples and the effort of speech-language pathologists can be effectively decreased. Moreover, the pronunciation confusion network is succeeded to improve the accuracy of automatic speech recognition. Using dependency network and multiattribute decision model, the likelihood for identification of articulation error pattern can be effectively estimated by integrating the information of testing phoneme, labeled results, and speech. Experimental results reveal the practicability of proposed method and system. It can be implemented and used in long-distance treatment and pre-screening.

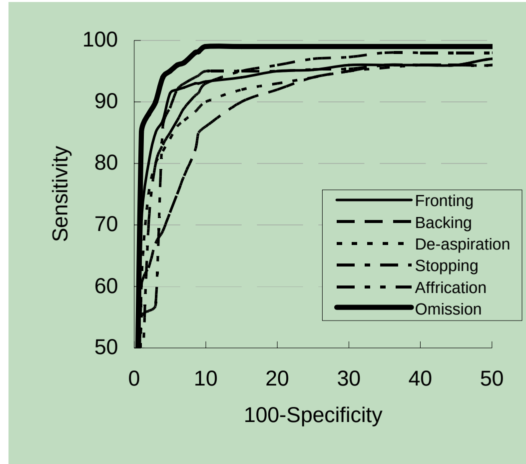

Fig. 7 The ROC curves for each identification results of articulation error patterns in training database

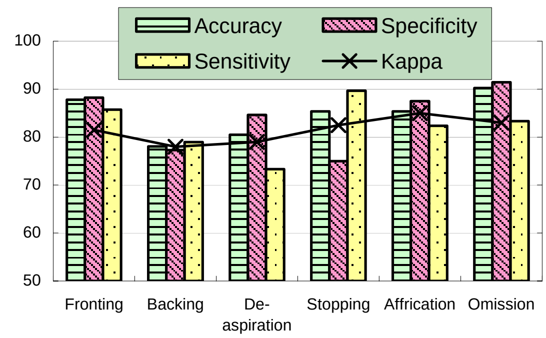

Fig. 8 The experimental results of identification of articulation error patterns

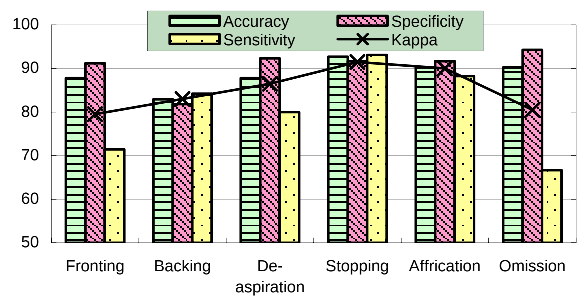

Fig. 9 The experimental results of identification of articulation error patterns with manual labeling

## REFERENCES


[1] J. E. Bernthal and W. B. Bankson, *Articulation and Phonological Disorders*. Allyn and Bacon, 2004.

[2] M.S. Shah and P.C. Pandey, "Estimation of Vocal Tract Shape for VCV Syllables for a Speech Training Aid," in Proc. IEEE International Conference of the Engineering in Medicine and Biology Society, pp. 6642-6645, 2005.

[3] J. Mullen, D.M. Howard, and D.T. Murphy, "Real-Time Dynamic Articulations in the 2-D Waveguide Mesh Vocal Tract Model," IEEE Trans. Audio, Speech and Language Processing, vol. 15, issue 2, pp.577-585, Feb. 2007.

[4] G. Georgoulas, V. C. Georgopoulos, and C. D. Stylios, "Speech Sound Classification and Detection of Articulation Disorders with Support Vector Machines and Wavelets," in Proc. the 28th IEEE International Conference on EMBS Annual, 2006.

[5] A. M. Stearns and S. A. Frisch, "Production and perception of place of articulation errors," Journal of Acoustic Society of America, vol. 120, no. 5, pp. 3251, Nov. 2006.

[6] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, and M. Stone, "Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 1245-1248, April 2007.

[7] F. Robineau, F. Boy, J.P. Orliaguet, J. Demongeot, and Y. Payan, "Guiding the Surgical Gesture Using an Electro-Tactile Stimulus Array on the Tongue: A Feasibility Study," IEEE Trans. Biomedical Engineering, vol. 54, issue 4, pp. 711-717, 2007.

[8] T. Orzechowski, A. Izworski, R. Tadeusiewicz, K. Chmurzynska, P. Radkowski, and I. Gatkowska, "Processing of pathological changes in speech caused by dysarthria," in Proc. of 2005 International

Symposium on Intelligent Signal Processing and Communication Systems, pp. 49-52, Dec. 2005.

[9] Gustavo P., Joao F. M., Jose H. S., and Alexandre L. M. L., “Voice Command Recognition with Dynamic Time Warping (DTW) using Graphics Processing Units (GPU) with Compute Unified Device Architecture (CUDA),” 19th International Symposium on Computer Architecture and High Performance Computing, pp. 19-25, Oct. 2007.

[10] 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. of IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, pp. 869-872, April 2007.

[11] M. Rajamanohar and E. Fosler-Lussier, “An evaluation of hierarchical articulatory feature detectors,” in Proc. of 2005 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 59-64, Nov. 2005.

[12] Y. Tian, Q. Yang, T. Huang, C.X. Ling, and W. Gao, “Learning Contextual Dependency Network Models for Link-Based Classification,” IEEE Trans. Knowledge and Data Engineering, vol. 18, issue 11, pp. 1482-1496, Nov. 2006.

[13] C. Preisach and L. Schmidt-Thieme, “Relational Ensemble Classification,” in Proc. Sixth International Conference on Data Mining, pp. 499-509, Dec. 2006.