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/ would be heard as /k4/.

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.
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 \( w_j = s_{1j}s_{2j}\cdots s_{nj} \), the \( s_{jm} \) is the \( m \)-th phoneme and modeled by Hidden Markov Model (HMM) [1]. The corresponding speech, \( o_j \) is used to automatically segment and label by maximum posterior probability as follows:

\[
\begin{align*}
\hat{s}_{jm} &= \arg \max_{s_{jm}} P(s_{jm} | o_j) \\
&= \arg \max_{s_{jm}} \left( \prod_{k=1}^{n} P(s_{jk} | s_{jm}) \right) \left( \prod_{k=1}^{n} P(o_k | s_{jk}) \right) \left( P(o_j | s_{jm}) \right) \left( P(s_{jm}) \right) \left( P(o_j) \right) \left( \frac{1}{M} \right) \\
&= \text{arg} \max_{s_{jm} \in \hat{s}_{jm}} \left( \prod_{k=1}^{n} P(s_{jk} | s_{jm}) \right) \left( \prod_{k=1}^{n} P(o_k | s_{jk}) \right) \left( P(o_j | s_{jm}) \right) \left( P(s_{jm}) \right) \left( P(o_j) \right) \left( \frac{1}{M} \right) \\
&= \text{arg} \max_{s_{jm} \in \hat{s}_{jm}} \left( \prod_{k=1}^{n} P(s_{jk} | s_{jm}) \right) \left( \prod_{k=1}^{n} P(o_k | s_{jk}) \right) \left( P(o_j | s_{jm}) \right) \left( P(s_{jm}) \right) \left( P(o_j) \right) \left( \frac{1}{M} \right)
\end{align*}
\]

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

\[
P(\hat{s}_{jm} | s_{jm}) = \frac{C(\hat{s}_{jm})}{C(s_{jm})}
\]

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 \( P(\hat{s}_{jm} | s_{jm}) \). \( \hat{s}_{jm} \) 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:

\[
L(E_i) = \left( \frac{1}{M} \sum_{w} P(E_i | s_{jm}, \hat{s}_{jm}) \right)^{1/\eta}
\]

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 \( \hat{s}_{jm} \) and language information \( \hat{s}_{jm} \) 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

\[
\begin{align*}
P(E_s w, \hat{s}_{jm}) &= P(E_s s_{jm} \hat{s}_{jm}) P(\hat{s}_{jm} | s_{jm}) P(o_{jm}) P(s_{jm}) \\
&\equiv P(E_s s_{jm} \hat{s}_{jm}) P(\hat{s}_{jm} | s_{jm}) P(o_{jm}) P(s_{jm}) \\
&\equiv P(E_s s_{jm} \hat{s}_{jm}) P(\hat{s}_{jm} | s_{jm}) P(o_{jm}) P(s_{jm}) \\
&\equiv P(E_s s_{jm} \hat{s}_{jm}) P(\hat{s}_{jm} | s_{jm}) P(o_{jm}) P(s_{jm})
\end{align*}
\]

Since the \( P(s_{jm}) \) and \( P(o_{jm}) \) 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

\[
P(E_s s_{jm} \hat{s}_{jm}) P(\hat{s}_{jm} | s_{jm}) P(o_{jm}) P(s_{jm})
\]

It is clearly that the probability \( P(\hat{s}_{jm} | s_{jm}) \) and \( P(o_{jm} | \hat{s}_{jm}) \) can be computed in the process of segmentation and labeling by automatic speech recognition. \( P(E' s_{jm} \hat{s}_{jm}) \) can be estimated
by MLE as

$$P(E_i|s_m) = \frac{C(E_i)}{C(s_m)}$$  \hspace{1cm} (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

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}.$$  \hspace{1cm} (7)$$

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

$$\text{Specificity} = \frac{TN}{(TN + FP)}.$$  \hspace{1cm} (8)$$

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

$$\text{Sensitivity} = \frac{TP}{(TP + FN)}.$$  \hspace{1cm} (9)$$

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

$$\text{Kappa} = \frac{(P_o - P_e)}{(1 - P_e)}.$$  \hspace{1cm} (10)$$

where

$$P_o = \frac{TP + TN}{(TP + TN + FP + FN)}$$  \hspace{1cm} (11)$$

and

$$P_e = \frac{(TP + FN)(TP + FP) + (FN + TN)(FP + TN)}{(TP + TN + FP + FN)}.$$  \hspace{1cm} (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
The likelihood for accuracy of automatic speech recognition. Using dependency pronunciation confusion network is succeeded to improve the pathologists can be effectively decreased. Moreover, the articulation examples and the effort of speech-language recognition is applied to automatic segment and label the captured friendly by photo naming task. Automatic speech examples of an individual’s articulation pattern could be the articulation error patterns of articulation disorders. The pre-screening.

Thus, choosing a suitable threshold, it can achieve high accuracy of identification of articulation error pattern was degraded by the error labeling of automatic speech language pathologists to preliminarily examine articulatory ability. Experimental results reveal the practicability of proposed method and system. It can be labeled results, and speech. 

<table>
<thead>
<tr>
<th>Error Pattern</th>
<th>Fronting</th>
<th>Backing</th>
<th>De-aspiration</th>
</tr>
</thead>
<tbody>
<tr>
<td>g→d</td>
<td>83.33</td>
<td>87.96</td>
<td>73.99</td>
</tr>
<tr>
<td>k→d</td>
<td>76.92</td>
<td>92.57</td>
<td>68.79</td>
</tr>
<tr>
<td>k→t</td>
<td>81.08</td>
<td>16.94</td>
<td>7.41</td>
</tr>
<tr>
<td>zhi→d</td>
<td>18.80</td>
<td>93.09</td>
<td>65.96</td>
</tr>
<tr>
<td>chi→d</td>
<td>27.48</td>
<td>10.42</td>
<td>76.92</td>
</tr>
<tr>
<td>chi→t</td>
<td>15.85</td>
<td>14.89</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stopping</th>
<th>Affrication</th>
<th>Omission</th>
</tr>
</thead>
<tbody>
<tr>
<td>f→b</td>
<td>75.26</td>
<td>66.15</td>
</tr>
<tr>
<td>l→g</td>
<td>90.91</td>
<td>81.58</td>
</tr>
<tr>
<td>ri→g</td>
<td>83.87</td>
<td>70.97</td>
</tr>
<tr>
<td>zi→d</td>
<td>28.13</td>
<td>75.00</td>
</tr>
<tr>
<td>zi→t</td>
<td>89.58</td>
<td>23.26</td>
</tr>
</tbody>
</table>

**TABLE I**

<table>
<thead>
<tr>
<th>Probability Distributions of Pronunciation Error for Each Articulation Error Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE (Pronunciation Error)</td>
</tr>
<tr>
<td>--------------------------</td>
</tr>
<tr>
<td>g→d</td>
</tr>
<tr>
<td>k→d</td>
</tr>
<tr>
<td>k→t</td>
</tr>
<tr>
<td>zhi→d</td>
</tr>
<tr>
<td>chi→d</td>
</tr>
<tr>
<td>chi→t</td>
</tr>
</tbody>
</table>

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


