AUTOMATIC ASSESSMENT OF ARTICULATION DISORDERS USING CONFIDENT UNIT-BASED MODEL ADAPTATION

Hung-Yu Su, Chun-Hsien Wu, and Pei-Jen Tsai

Department of Computer Science and Information Engineering,
National Cheng-Kung University, Tainan, Taiwan, R.O.C.

ABSTRACT

This paper presents an approach to automatic assessment on articulation disorders using unsupervised acoustic model adaptation. Prior knowledge is obtained via the phonological analysis of the speech data from 453 articulation disordered children. A confusion matrix of the recognition units for a specific subject is re-estimated based on the prior knowledge and the recognition results to choose the confident units for adaptation. The adapted acoustic models can effectively improve the recognition performance of the disordered speech and thus used for articulation assessment. In the experiments, the proposed unsupervised adaptation method achieved a significant performance improvement of 9.1% for disordered speech on syllable recognition rate. Automatic assessment also shows encouraging consistency to the assessment from the therapist.

Index Terms — Articulation disorder, articulation assessment, speaker adaptation, speech recognition

1. INTRODUCTION

Articulation disorders (ADs) [1] caused by disabilities on articulators or dysarthria bring barriers on speech communication, and therefore influences the education, sociality and daily life. Speech therapists provide articulation assessment and training to rectify the subjects with ADs. The AD population is about 2.5 million with different AD types in Taiwan [2], but only about 300 speech therapists work in hospitals to provide services for articulation disordered people. Augmentative and Alternative Communication (AAC) technologies provide computer-assisted services to lessen the effects of hindered understanding for AD speakers. To design an automatic process with AAC technologies for articulation assessment is helpful for therapists and AD people.

Speaker adaptation is the most popular technique to improve recognition performance for a specific speaker. Maximum Likelihood Linear Regression (MLLR) [3] and Maximum A Posteriori (MAP) [4] are the mostly used approaches in speaker adaptation. Several works on dysarthric speech [5]-[7] had been addressed. For environment control using automatic speech recognizer (ASR) [8][9] have been proposed recently. ASR is trained or adapted from the general models using the speech of a specific AD speaker for keyword recognition.

Two problems for modeling the AD speech are 1) insufficient training database and un-predictable variations from AD speakers generally degrades the recognition performance and 2) traditional speaker adaptation adapts the acoustic models to overfit the incorrect pronunciation while miss the purpose for articulation disorder assessment. In the assessment of ADs, the purpose is to identify the mispronounced phones from the speech pronounced by the AD speaker. Instead of adapting the acoustic models to fit the articulation disordered speech, this study tries to adapt the acoustic models using the correctly pronounced speech from the AD speaker while keep the incorrectly pronounced units unchanged. Finally, the speaker-dependent acoustic models can be applied for articulation disorder assessment of the AD speaker.

Strategies of speaker adaptation can be divided into supervised/unsupervised and incremental/batch. In this paper, an unsupervised, incremental adaptation is proposed to adapt the acoustic models for AD speakers from the acoustic models of normal speakers. Manual articulation disorder assessments of 453 AD children were conducted to obtain the mispronounced phenomenon between phones. This information is adopted to re-estimate the confusion matrix of the phones for the AD speaker. The re-estimation process keeps the confident/consistent recognition results of the speech pronounced by the AD speaker for adaptation. Incrementally, the adapted acoustic models recognize the speech of the AD speaker and obtain more confident recognition results. The process is ended until no more new phones confirmed for adaptation. The final re-estimated confusion matrix provides the mispronunciations of the phones for the AD speakers and the results can be analyzed based on the pre-defined articulation and phonological error types to obtain the assessment results.

2. ANALYSIS OF ARTICULATION DISORDERS

To achieve the aim of this investigation, phonological information is considered as the prior knowledge to choose the correctly pronounced and recognized phones for adaptation. This section introduces the characteristics of the phones pronounced by AD speakers.

2.1. Articulation disorder from phonological analysis

Articulation is a series of interactions among articulators and breath. AD occurs for varied reasons, e.g. hearing loss, oral-motor problems, or dysarthria. In [1] and [10], the pronunciation errors can be summarized as the following four types:

- Substitution: a phone is substituted by another one.
- Omission: a phone is omitted.
- Distortion: a phone is distorted with a non-standard one.
- Addition: a phone is added to the pronounced phone.

The error types based on phonological analysis are classified according to the articulation manners and places [1][11] for further information in assessing the AD speakers. Here are some examples of the 14 error types used by therapists:

- Fronting: replacing alveolar by velar.
- Backing: replacing velar by alveolar.
- Aspirating: replacing un-aspirated by aspirated.
- Stopping: replacing by stops.
- Affricating: replacing by affricates.
- Etc.

These definitions can be used to indicate the specific error type on articulation of a speaker from the confusion matrix of the phones. According to these error types, therapist can provide proper training courses for the AD subjects.

2.2. Prior knowledge of articulation disorders

The assessments on 453 AD children were collected by the Department of Otolaryngology in National Cheng Kung University Hospital. In these assessments, 16,314 tested syllables were gathered from subjects. There are 5,969 substitutions, 1,246 deletions, 15 distortions, and 2 insertions for consonants. For vowels, the occurrences of AD are too rare to be considered in this study.

3. UNSUPERVISED ADAPTATION USING ARTICULATION DISORDERED SPEECH

In speaker adaptation, the transcription of a phone is used to determine which model should be adapted. For the description of model adaptation, the base form is defined as the canonical pronunciation of the phone, while the surface form represents the observed (actual) pronunciation. For the speech from an AD speaker, the surface forms do not match the base forms due to the unmatched acoustic models and

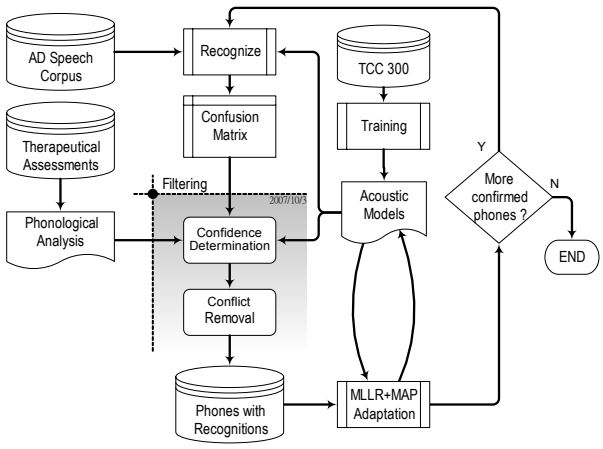


Figure 1 The architecture of unsupervised adaptation for AD speech

articulation disorders. To deal with the above two problems for an AD speaker, an unsupervised adaptation mechanism based on confident units is adopted. Figure 1 illustrates the framework of the proposed approach.

3.1. Confident unit selection from the recognized results

In the preliminary study, the speech of the AD speaker is recorded by reading a set of sentences balanced on 408 non-tonal Mandarin syllables (each syllable appears 7-8 times). A speaker-independent ASR for normal speech, trained from TCC300 corpus [12], is used. Starting with the prototype ASR, the syllable recognition rate of AD speech is getting worse due to the errors in pronunciations (only 27.4 %). A consonant-based confusion matrix with the elements being the probabilities of the recognized results from the base form is constructed. To determine which mispronunciations are pronounced by the speaker and recognized confidently, a process of re-estimation is proposed.

3.1.1. Confidence determination

To determine the confident elements in the confusion matrix, the probability of the recognized phone r given the base form b is computed as:

$$\hat{P}(r \mid b) = \sum_{S} P(r, s \mid b) \approx \sum_{S} P(r \mid s) P(s \mid b) \quad (1)$$

where s indicates the surface forms produced for b by the AD subject in the speech corpus for assessment. $P(s \mid b)$ is the probability for pronouncing the b as s, and the distribution is obtained from the assessment results from the 453 AD children. $P(r \mid s)$ is the probability for recognizing surface form s by the acoustic model r and estimated:

$$P(r \mid s) = \frac{1}{N} \sum_{n} \frac{A_r(s_n)}{\sum_{r} A_r(s_n)} \quad (2)$$

where $A_r(s)$ indicates the likelihood of *s* recognized by the acoustic model *r*. *N* is the number of the phones pronounced as *s* by the subject. If probability $P(r \mid b)$ is greater than $\hat{P}(r \mid b)$, the phones pronounced for *b* and recognized as *r* in the confusion matrix are kept for adapting the acoustic model *r*.

3.1.2. Conflict removal

According to the investigation of the AD speech, the ADs usually happen uni-directionally. That is, for an AD speaker, if phone /*b*/ is mispronounced as /*p*/, the speaker is unlikely to mispronounce /*p*/ backward to /*b*/. For symmetric elements in the confusion matrix, the elements with probabilities smaller than a threshold will be removed. Figure 2 illustrates the confusion matrix for the original ASR and the confusion matrix for the adapted ASR. The gray level of the element indicates the probability of each element, and the segmented phones with recognition results in these elements are used to adapt the acoustic models.

3.2. Incremental adaptation

After the re-estimation process, the confident phones with recognition results are adopted for adaptation. The adapted acoustic models can perform better recognition of the speech from the assessed subject. In this investigation, incremental adaptation mechanism based on this assumption is proposed to achieve the aim of unsupervised adaptation, and terminated while the number of confirmed phones for adaptation does not increase from the previous iteration.

MLLR and MAP algorithms are used to adapt the general acoustic models by the collected AD speech. MLLR adapts the means of the Gaussian mixtures in the models by:

$$\hat{\mu} = W\xi \quad (3)$$

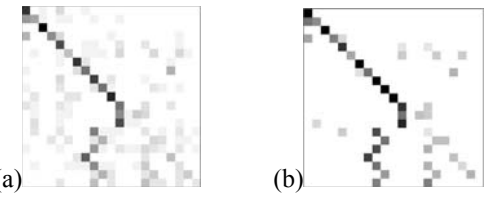
where $\hat{\mu}$ is the vector of the adapted means of the mixtures in an HMM state. *W* is the $n \times (n+1)$ transformation matrix for the state, and *n* is the number of mixtures in a state. ξ is an extended mean vector defined as:

$$\xi = [\omega, \mu_1, \mu_2 ... \mu_n] = [\omega : \mu]' \quad (4)$$

where μ is the original mean vector of the state and ω is an offset term for distinguishing the training data and adaptation data. The transformation matrix $\hat{W}$ is determined by maximizing the equation with observed data set θ:

$$\hat{W} = \arg\max_{\theta} P(\theta \mid W) \quad (5)$$

MAP adapts the means $\hat{\mu}_{jm}$ of the j_{th} mixtures in the m_{th} state by:



(a) (b)

Figure 2 (a) Original confusion matrix from ASR and (b) re-estimated confusion matrix for AD subject

Table 1 Recognition rate of ASR on syllable level with consideration of right context dependent (RCD) information

		Original	RCD removed
Normal speech	Inside	82.3 %	77.5 %
	Outside	62.4 %	52.1 %
AD speech		27.5 %	29.2 %

Table 2 Recognition rates of the proposed unsupervised adaptation on syllable level (inside)

Adaptation Data	Base-form	Proposed Adaptation	Surface-form
Recognition rate	62.3 %	68.19 %	79.3 %

Table 3 Precision and recall rates of automatic assessment compared to the results from therapist

	Precision	Recall
Mispronounced consonants	100 %	73.3 %
Phonological errors	75 %	81.8 %

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm} + \tau} \bar{\mu}_{jm} + \frac{\tau}{N_{jm} + \tau} \mu_{jm} \quad (6)$$

where $\bar{\mu}_{jm}$ and μ_{jm} are the means of the adaptation data and the mixture of the original state model, respectively. N_{jm} indicates the number of adaptation data and τ is the ratio of the adaptation data. Combining Eq. (3) and Eq. (6) in adaptation provides more precise acoustic models.

4. EXPERIMENTS

To evaluate the proposed approach, a college student with articulation disorder was invited to participate in the experiment. In evaluating the performance of automatic assessment, the subject was firstly assessed by a certificated speech therapist and the manual assessment was regarded as a gold standard for comparison.

4.1. Experiment environment

In this investigation, the ASR for general speakers is trained by the TCC300 speech corpus with 151 acoustic sub-syllable models (112 for context-dependent consonants, 38 for context-independent vowels, and 1 for silence). There are 3 states for consonant and 4 states for vowel, and each state is modeled by 16 Gaussian models. 39 acoustic features were used: 12 MFCCs, 12 ΔMFCCs, 12 $\Delta\Delta$

MFCCs, 1 log energy, 1 Δ log energy, and 1 $\Delta\Delta$ log energy.

The test corpus of AD speech was recorded from a subject with hearing loss of 60 dB and equipped with cochlear implant. 112 sentences balanced on 408 non-tonal Mandarin syllables were recorded by the subject. Except for the base forms of these 112 sentences, the surface forms of the corpus were annotated by a speech therapist for evaluation.

4.2. Evaluation on ASR

Before examining the performance of the proposed approach, a brief evaluation for ASR is provided in Table 1. The recognition rates for normal speech are 82.3% and 62.4% for inside and outside tests, respectively. For AD speech recognition, the recognition rate achieves only 27.5%. Right-Context-Dependent (RCD) criterion is used in speech recognition to constrain the combinations of sub-syllables. In AD speech, many combinations unsatisfying the RCD constraint occurs. The removal of the RCD decreases the recognition rate of normal speech; while slightly improve the performance on AD speech.

4.2. Performances of adaptations

The original ASR is adapted using two transcriptions: base form of the AD speech corpus as the base-line and the other is the surface form of the speech annotated manually as the best case. Table 2 shows the syllable recognition rates of ASRs adapted by the base forms, surface forms and the proposed approach. The system adapted with surface forms achieved better performance than the base-line ASR on recognizing the surface form. The proposed adaptation method provides a result better than the base-line ASR.

4.3. The performance of automatic assessment

For the assessment of articulation disorder, the error types defined from phonological analysis can be obtained from the re-estimated confusion matrix. Table 3 shows the comparison for automatic assessment and the gold standard. For the mispronounced consonants, the precision and recall rates are 100% and 73.3%, respectively. For phonological analysis of errors, therapist lists 12 consonants distributed in four error types while the proposed approach concluded 11 consonants in five error types. The error type precision and recall rates are 75% and 81.8%, respectively.

5. CONCLUSION

This paper presents an unsupervised acoustic model adaptation for AD speakers. The mechanism incrementally adapts acoustic models with confirmed surface form from the prior knowledge and the likelihood of acoustic models for the speech from the AD speakers. Distorted or misrecognized phones can be gradually recognized stably in the iterative re-estimation procedure. The adapted models can recognize the AD speech well only in the sub-syllables, and the results contain the articulation disorder information of the speaker and ASR errors. The information is valuable in helping the therapist design the articulation training courses for the AD speakers, but the errors of ASR need further analysis to eliminate. On the other hand, the stable elements in the confusion matrix can be used to repair the syllable sequence to obtain the correct sentence the AD speaker intends to present.

6. ACKOWLEDGEMENT

The authors would like to thank the Dr. Jiunn-Liang Wu and Yi-Hui Lin in the Department of Otolaryngology in National Cheng Kung University Hospital for providing great assistance on this study.

7. REFERENCES

[1]. B. G. Lin, *Articulation Disorder and Treatment*, Wunan Book co., ltd. 1994.

[2]. Available: Http://www.moi.gov.tw/stat/english/index.asp.

[3]. C. Leggetter and P. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMM", *Computer Speech and Language*, 9: pp. 1490-1499, 1995.

[4]. G. Zavagliakos, R. Schwartz and J. McDonough, "Maximum a Posteriori adaptation for Large Scale HMM Recognizers", in *Int. Conf. Acoustics, Speech, Signal Processing* '96, Atlanta GA, pp. 725-728, 1996.

[5]. B. Blaney and J. Wilson, "Acoustic Variability in Dysarthria and Computer Speech Recognition", *Clinical Linguistics and Phonetics*, 14(4): pp 307-327, 2000.

[6]. D. Bowes, "Getting It Right and Making It Work! Selecting the Right Speech Input and Writing Software for Users with Special Needs", *Proc. of Technology and Persons with Disabilities*, California, State University Northridge, 1999.

[7]. P. Doyle, H. Leeper, A-L. Kotler, and N. Thomas-Stonell, "Dysarthric Speech, a Comparison of Computerized Speech Recognition and Listener Intelligibility", *Journal of Rehabilitation Research and Development*, 34(3): pp 309-316, 1997.

[8]. P. Green, J. Carmichael, A. Harzis, P. Enderby, M. Hawley and M. Parker, "Automatic Speech Recognition with Sparse Training Data for Dysarthric Speakers", *Eurospeech*, pp. 1189–1192, 2003.

[9]. H. Matsumasa, K. Tanaka, T. Takiguchi, Y. Ariki, I. Li, and T. Nakabayashi, "Evaluation of Speech Recognition by a Person with Articulation Disorder in Operation for Home Information Applications", *IEICE technical report. Welfare Information technology*, Vol.107, No.61, pp. 33-38, 2007.

[10]. X. J. Lai, "Articulation disorder", The Magazine of Hearing and Language, The Speech-Language-Hearing Association of Republic of China, Vol. 4, pp 70-73 1987.

[11]. X. J. Lai, "*Diagnostic methods of speech and management*," the Magazine of Language Therapy, Taipei city Government Educational site, pp. 123-133, 1990.

[12]. Available: Http://www.aclclp.org.tw/use_mat.php#tcc300edu