An Articulation Training System with Intelligent Interface and Multimode Feedbacks to Articulation Disorders

Yeou-Jiunn Chen
Department of Electrical Engineering, Southern Taiwan University, Tainan County, Taiwan
chenyj@mail.stut.edu.tw

Jiunn-Liang Wu, Hui-Mei Yang
Department of Otolaryngology, National Cheng Kung University Hospital, Tainan, Taiwan
jiunn@mail.ncku.edu.tw

Chung-Hsien Wu
Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan
chwu@csie.ncku.edu.tw

Chih-Chang Chen, Shan-Shan Ju
Human-Interaction Technology Center, Industrial Technology Research Institute, Tainan County, Taiwan
{nickchen, ssju}@itri.org.tw

Abstract

Articulation training with many kinds of stimulus and messages such as visual, voice, and articulatory information can teach user to pronounce correctly and improve user's articulatory ability. In this paper, an articulation training system with intelligent interface and multimode feedbacks is proposed to improve the performance of articulation training. Clinical knowledge of speech evaluation is used to design the dependent network. Then, automatic speech recognition with dependent network is applied to identify the pronunciation errors. Besides, hierarchical Bayesian network is proposed to recognize user's emotion from speeches. With the information of pronunciation errors and user's emotional state, the articulation training sentences can be dynamically selected. Finally, a 3D facial animation is provided to teach users to pronounce a sentence by using speech, lip motion, and tongue motion. Experimental results reveal the usefulness of proposed method and system.

1. Introduction

Articulation errors generate different degrees of abnormality in articulation and seriously reduce speech intelligibility and the ease of spoken communication. The population of language disorder aged from 4 years old to 15 years old is 2.64 percent, and amount them the population of articulation disorder is 43.36 percent in Taiwan. Moreover, the speech-language pathologists worked in hospital are insufficient. Traditionally, a speech-language pathologist uses his or her clinical experience to identify articulation error patterns and provide suitable training courses, which are time-consuming and expensive processes. Therefore, developing a computer-assisted system for articulation training is very helpful for speech-language pathologists and articulation disorders.

Speech-language pathologists classify the articulation into six articulation error patterns: fronting, backing, de-aspiration, stopping, affrication, and omission [1]. Those articulation error patterns are examined in terms of the place of articulation and the manner of articulation. Each articulation error pattern will generate several pronunciation errors. In articulation training, the pronunciation errors should be found and used to teach user's pronunciation. 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/. In this case, the pronunciation error is /k/→/t/.

Recently, automatic speech recognition (ASR) had been widely applied to many applications. ASR can be used to identify the articulation attributes and it is very useful to automatically identify the pronunciation errors [2]. In articulatory training materials, a phoneme will appear in several sentences and the effect of recognition errors generated from ASR can be reduced. Besides, dependency network (DN) had been applied to collective classification and knowledge discovery [3, 4]. It is well suited to the task of predicting preferences and is generally useful for probabilistic inference. From the clinical practice and experience, the graphical representation of DN can be easily designed to represent the relationships of speech, training sentences, and pronunciation error. Thus, it is appropriate to integrate clinical experience into identification of pronunciation errors.

For articulation training, researches usually aimed at computer-assisted treatment by using visual feedback [4]. However, visual feedback provides acoustic and articulatory parameters, such as speech intensity, spectrogram, pitch, formants, vocal tract shape, and lip shape. Those feedbacks are very difficult for users to map the visual feedbacks to the manner of articulation and the place of articulation. Many researches had also provided facial animations [5] and most of them only

focus on facial expression and lip movement. Therefore, a feedback with tongue movement can effectively provide articulatory cues to help articulation disorders in articulation training.

Traditionally, emotions are classified into many categories: fear, anger, joy, sadness and disgust [6]. For a computer-assisted system, it is very important to promote the interest in practicing pronunciation of articulation training. Thus, many researches had proposed emotion recognition to enhance human-machine interface [7]. People can recognize emotional speeches with about 60% accuracy and some emotion recognition systems are 50-60%. Generally, there are three emotions, named positive, negative, and neutral, in human-machine interface. It is very useful to detect user's emotional state, then, select suitable sentence of articulation training. Moreover, the system may be end in several turns to keep user's interest.

In this study, we propose an articulation training system with intelligent interface and multimode feedbacks to help articulation disorders, see Figure 1. First, a sentence with multimode feedbacks is generated to help user to simulate correct pronunciation. Second, the user is asked to pronounce this sentence. According to the pronounced speech, the pronunciation errors and emotional state of a user are automatically identified by ASR with DN and hierarchical Bayesian network (HBN). Finally, the system can select several suitable sentences from articulatory training materials or may end in several turns.

2. Identification of Pronunciation Errors

When a user is asked to pronounce a sentence with N_i phonemes, $w_i = s_1 s_2 \cdots s_{N_i}$, the corresponding speech observations O_i is recorded. We use ASR with DN to segment the speech into a sequence of speech segments, $O_i = o_1 o_2 \cdots o_{N_i}$, and labeled phonemes, $\hat{w}_i = \hat{s}_1 \hat{s}_2 \cdots \hat{s}_{N_i}$. With ASR, the speech segments and labeled phonemes with maximum posterior probability as follows:

$$\begin{aligned} \hat{w}_i &= \arg\max_{\tilde{w}_i} P(\tilde{w}_i | O_i) \\ &= \arg\max_{\tilde{s}_1 \tilde{s}_2 \cdots \tilde{s}_{N_i}} P(\tilde{s}_1 s_1 o_1 \tilde{s}_2 s_2 o_2 \cdots \tilde{s}_{N_i} s_{N_i} o_{N_i}) . \end{aligned} \quad (1)$$

For an articulation disorder, the articulatory information of phonemes is consistent with each other. Therefore, each phoneme is assumed to be independent and Eq. (1) is rewritten as

$$\hat{w}_i = \arg\max_{\tilde{s}_1 \tilde{s}_2 \cdots \tilde{s}_{N_i}} \prod_{m=1}^{N_i} P(\tilde{s}_m s_m o_m) \quad (2)$$

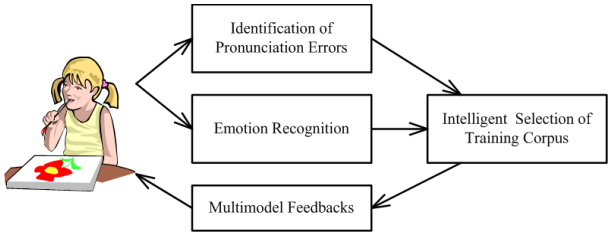

Figure 1. The block diagram of proposed articulation training system

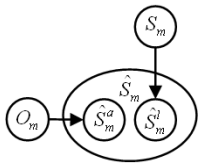

Figure 2. Dependency network for phoneme labeling

Since the labeling process of $\hat{s}_m$ contains linguistic and acoustic information, which are denoted by $\hat{s}_m^a$ and $\hat{s}_m^l$, the DN is designed as Figure 2. Hence, the joint probability of DN consists of a set of conditional probability distributions:

$$\begin{aligned} P(\hat{s}_m s_m o_m) &= P(\tilde{s}_m | s_m o_m) P(o_m) P(s_m) \\ &\cong \left(P(\tilde{s}_m^l | s_m)^{\omega_l} P(\tilde{s}_m^a | o_m)^{\omega_a} \right)^{1/\omega_l + \omega_a} \end{aligned} \quad (3)$$

where w_l and w_a are the weighting factors for language and acoustic information. $P(\tilde{s}_m^l | s_m)$ and $P(\hat{s}_m^a | o_m)$ are the probabilities generated from language and acoustic information. Finally, the probability of labeling result can be estimated as

$$\hat{w}_i = \arg\max_{\tilde{s}_1 \tilde{s}_2 \cdots \tilde{s}_{N_i}} \prod_{j=1}^{N_i} \left(P(\tilde{s}_j^l | s_j)^{\omega_l} P(\tilde{s}_j^a | o_j)^{\omega_a} \right)^{1/\omega_l + \omega_a} . \quad (4)$$

In this study, $P(\tilde{s}_j^a | o_j)$ is estimated by hidden Markov model [8] and $P(\tilde{s}_j^l | s_j)$ is estimated by maximum likelihood estimation as

$$P(\tilde{s}_j | s_j) = \frac{P(\tilde{s}_j s_j)}{P(s_j)} = \frac{C(\tilde{s}_j s_j)}{C(s_j)} \quad (5)$$

where $C(\bullet)$ is the number of instances appeared in training corpus.

3. Emotion Recognition

The features in emotional speech can be categorized as frame-based and sentence-based features. The frame-based features contain cepstrum and mel-frequency ceptral coefficient. The sentence-based features are pitch and intensity features such as short-

term energy maximum, variance of short-term energy, F0 maximum, and position of F0 maximum [18].

HBN allows the random variables to represent arbitrarily structured types. Within a single node, there may also be links between components, representing probabilistic dependencies among parts of the structure. A HBN (see Figure 3) consists of structural part and conditional probability part. The structural part described the relationships between nodes and the conditional probability can quantify the links of structural part.

In this study, the frame-based and sentence-based features are modeled at level 1 and level 2 of HBN, respectively. Therefore, given a set of N variables $H_{1:N} = H_1, \ldots, H_N$, the joint probability distribution $P(H_{1:N}) = P(H_1, \ldots, H_N)$ can be factored into a sparse set of conditional probabilities as follows according to the conditional independency:

$$P(H_{1:N}) = \prod_{i=1}^{N} P\left(H_i \left| pa(H_i)\right.\right) \tag{6}$$

where $pa(H_i)$ is the set of parent nodes of node H_i in HBN.

4. Multimode Feedbacks

Multimode feedbacks including speech, lip motion, and tongue motion are generated by text-to-speech, speech segmentation, and 3D facial animation. First, text-to-speech is used to generate speech signal and corresponding boundaries of syllables. Second, sequential forward selection is used to find the boundaries. The phonemes can be categorized as consonant (C) and vowel (V). Hotelling's T^2 test statistic and Bayesian information criterion (BIC) are used to find the CV (C following V or V following C) and VV (V following V) boundaries, respectively.

Using likelihood-ratio procedure approach, the Hotelling's T^2 test statistic can be written as $T_b^2 = y_b' \Sigma_b^{-1} y_b$ where $y_b = \sqrt{b(N-b)/N}\,(\mu_1 - \mu_2)$. μ_1 and μ_2 is the mean of speech segments before and after boundary b, respectively. Σ is the common covariance matrix of speech observations and N is the number of features in speech observation. BIC is written as

$$\begin{aligned}\Delta BIC(b) = &\frac{1}{2}\left(N\log|\Sigma| - b\log|\Sigma_1| - (N-b)\log|\Sigma_2|\right) \\ &-\frac{1}{2}\lambda\left(d + \frac{1}{2}d(d+1)\right)\log N\end{aligned} \tag{7}$$

where Σ_1 and Σ_2 are the variance of segments before and after boundary b. λ is the penalty factor to compensate for small sample size cases.

Base on knowledge of acoustic phonetics, lip motions and tongue motions are defined to represent the motions of all 408 Mandarin syllables. Given a feature point with location (x_t, y_t) in frame t, the location of this feature point in frame t+1 is decided by:

$$(x_{t+1}, y_{t+1}) = \underset{(x_{t+1}, y_{t+1})}{\arg\min}\left(\sum_{i=\frac{-m}{2}}^{\frac{m}{2}}\sum_{j=\frac{-m}{2}}^{\frac{m}{2}}\left(\begin{matrix} I(x_t + i, y_t + j) - \\ I(x_t + i + u, y_t + j + v)^2 \end{matrix}\right)\right) \tag{8}$$

where $-z/2 \le u \le z/2$ and $-z/2 \le v \le z/2$. $I(x, y)$ indicates the intensity of pixel (x, y). z is the block size that contains the possible location of the feature point in frame $t+1$. Therefore, the control points are transformed to feature points of 3D facial models. Finally, the B-spline is applied to smooth the sequence of parametric points of 3D facial models.

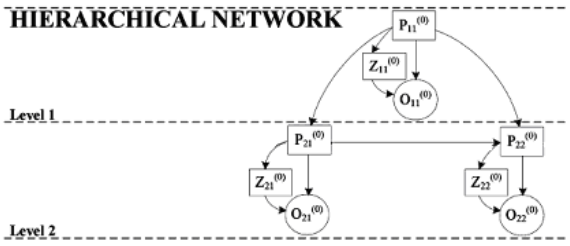

Figure 3. An example of hierarchical Bayesian network

5. Experimental Materials and Results

For identification of pronunciation errors, a Mandarin speech corpus named TCC300 was used to train the speaker-independent based ASR. TCC300 was recorded in ordinary office environments via four close-talking microphones. 14266 short and long sentences (about 16 hours) uttered by 100 males and 100 females were used for model training of automatic speech recognition. For training of DN, there were 553 (346 males and 207 females) articulation disorders, average age was 6 years, collected and the pronunciation errors were manually labeled by speech-language pathologists. We need to ignore pronunciation errors with low probability by select a suitable threshold. This value can be obtained by computing the histogram of probabilities of pronunciation errors for all phonemes. Therefore, all pronunciation errors with equal probability were treated as a baseline system and used to compare with DN. The correct rates of identification of pronunciation errors were 81.44% and 91.72% for baseline and DN, respectively. Furthermore, 370 sentences with three types of emotion (named positive, negative, and neutral) were collected to evaluate the performance of emotion recognition. The recognition model with Gaussian mixture model (GMM) was applied to compare the performance. The average recognition rates were 57.75% and 66.88% for GMM and HBN.

In this experiment, the spontaneous speech corpus with 3647 sentences was collected to evaluate the performance of phoneme segmentation. The results of phoneme segmentation were shown in Figure 4. The experimental results achieve practical performance. Besides, parts of synthesized 3D facial animation including speech signal, facial models, lip models, and tongue model were shown in Figure 5. The users give mean opinion scores (MOS) on a scale of 1 to 5, i.e., 5 for excellent level, 4 for good level, 3 for fair level, 2 for poor level, and 1 for unsatisfactory level. The average MOS was 4.07. Finally, the articulation training system with intelligent interface and multimode feedbacks was shown in Figure 6. The users were also asked to give MOS and the average MOS was 4.3. The users indicate that the training sentences are suitable to user's pronunciation errors and can satisfy user's emotion.

6. Conclusions

In this paper, we have presented an innovative approach to articulation training with intelligent interface and multimode feedbacks. We have used ASR with DN to automatically detect user's pronunciation errors. Using HBN, user's emotional state is effectively identified to dynamically design training materials. Multimode feedbacks helps articulation disorder to simulate the articulatory behavior. The experimental results have shown a practical implementation in performance. In the future, we would like to evaluate clinical performance of articulation training.

7. References

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

[2] S. M. Siniscalchi, P. Schwarz, and C.H. Lee, "High-Accuracy Phone Recognition By Combining High-Performance Lattice Generation and Knowledge Based Rescoring," *IEEE Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2007, pp. 869-872.

[3] C. Preisach and L. Schmidt-Thieme,). "Relational Ensemble Classification," Proceedings of *Sixth International Conference on Data Mining*, 2006, pp. 499-509.

[4] 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*, 18(11), 2006, pp. 1482-1496.

[5] Y. Sheng; A.H. Sadka, and A.M. Kondoz, "Automatic Single View-Based 3-D Face Synthesis for Unsupervised Multimedia Applications," *IEEE Transactions on Circuits and Systems for Video Technology*. 18(7), 2008, pp. 961-974.

[6] J.L. Arnott and I. Murray, "Towards the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion," *Jouranl of the Acoustic Society of America*, 1993, pp. 1097-1108.

[7] Y. Wang and L. Guan, "Recognizing Human Emotional State From Audiovisual Signals," *IEEE Transactions on Multimedia*, 10(4), 2008, pp. 659-668.

[8] H.T. Bunnel, D.M. Yarrington, and J.B. Polikoff, "Using Markov models to assess articulation errors in young children," *Journal of the Acoustical Society of America*, 107(5), 2000, pp. 2903.

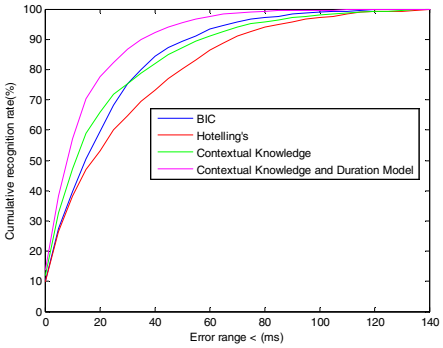

Figure 4. Results of phoneme segmentation

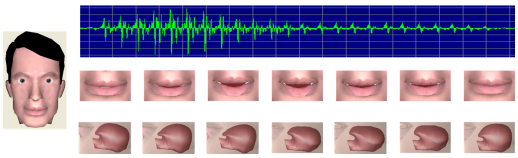

Figure 5. Parts of synthesized 3D facial animation for 'phong'

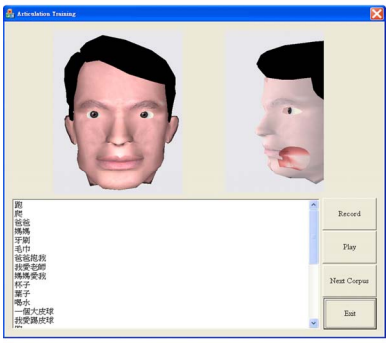

Figure 6. The intelligent interface of articulation training system

8. Acknowledgments

The authors would like to thank the National Science Council R.O.C. and Industrial Technology Research Institute for its financial support of this work, under Contract No. NSC 98-2221-E-218-020-MY2 and No. 120970183.