Development and Evaluation of On/Off Control for Electrolaryngeal Speech Via Artificial Neural Network Based on Visual Information of Lips

Liang Wu, Congying Wan, Supin Wang, and Mingxi Wan, *Xi'an, P. R. China*

Summary: Objective. To realize an accurate and automatic on/off control of electrolarynx (EL), an artificial neural network (ANN) was introduced for switch identification based on visual information of lips and implemented by an experimental system (ANN-EL). The objective was to confirm the feasibility of the ANN method and evaluate the performance of ANN-EL in Mandarin speech.

Study Design and Methods. Totally five volunteers (one laryngectomee and four normal speakers) participated in the whole process of experiments. First, trained ANN was tested to assess switch identification performance of ANN method. Then, voice initiation/termination time, speech fluency, and word intelligibility were measured and compared with button-EL and video-EL to evaluate on/off control performance of ANN-EL.

Results. The test showed that ANN method performed accurate switch identification (>99%). ANN-EL was as fast as normal voice and button-EL in onset control, but a little slower in offset control. ANN-EL could provide a fluent voice source with rare breaks (<1%) for a continuous speech. The results also indicated that on/off control performance of ANN-EL had a significant impact on perception, lowering the word intelligibility compared with button-EL. However, the words produced by ANN-EL were more intelligible than video-EL by approximately 20%.

Conclusions. The ANN method was proved feasible and effective for switch identification based on visual information of lips. The ANN-EL could provide an accurate on/off control for fluent Mandarin speech.

Key Words: Artificial neural network–Electrolarynx–On/off control–Visual information.

INTRODUCTION

Speech is the most important and efficient way of communication. Owing to laryngeal cancer or trauma, people are prone to remove their entire larynx and, therefore, lose their physiological structure for normal speech. However, taking advantage of the remaining vocal tract and principle of speech production, electrolarynx (EL) speech is an effective way for voice rehabilitation and alaryngeal communication.

The EL is a handheld and battery-powered device, which transmits mechanical vibration into laryngopharynx through the neck or into posterior oral cavity with a tube or denture.[1] Owing to easy learning and no additional surgery required, EL speech has been widely accepted for daily communication by more than one-half of the laryngectomees.[1–3] However, the conventional EL is not convenient for users. The occupation of one hand to hold EL and control an on/off button during speech is ranked in the top five deficits of EL communication.[4] Therefore, many methods have been reported on switch control for hands-free EL.

There are three representative methods used for on/off control of EL without hands: tongue, electromyography (EMG), and visual information (video) control. Knorr, Zwitman, and colleagues[5,6] designed a wireless intraoral EL, which was switched on/off by the tongue. Although hands were not required for holding the device and pushing the button, users still had to control on/off actively. Goldstein et al[7] found that it was feasible to control initiation and termination of EL voice automatically by EMG signals from neck strap muscles. EMG-EL realized the hands-free control and was easy to master after proper training.[8] However, the biggest disadvantage of EMG control was an additional surgery to preserve the omohyoid strap muscles.[9] This would result in more pain and expenditure to the patients. To address this problem, Stepp et al[10] used neck and face surface EMG (sEMG) to control onset and offset of EL, and found that individuals were able to use sEMG from multiple recording locations to produce running speech perceived as natural as that produced with a typical handheld EL.

Visual information, especially the shape information of lips, has been extensively used in speech recognition,[11,12] speaker identification,[13,14] and perceptual evaluation[15,16] because of its close relationship with speech production. Recently, a noncontact method based on lip deformation was proposed for automatic on/off control of an EL (video-EL) by Wan et al.[17] The shape of lip outer contour was extracted through real-time video processing and presented by an ellipse with two parameters, namely the semimajor (a) and semiminor axes (b). Finally, the ratio of b to a (b/a) was used to determine switch on/off through a single threshold judgment. Wan et al[17] reported that video-EL was effective in the automatic on/off control and could produce fluent Mandarin speech as intelligible as button-EL. However, owing to the single parameter (b/a) and single threshold used in b/a method, video-EL could not generate voice initiation and termination as fast as button-EL, which affected the perception of isolated word. First, only one parameter (b/a) is limited to represent and differentiate all the lip shapes of phonation from silence. Second, the

Accepted for publication October 22, 2012.

From the The Key Laboratory of Biomedical Information Engineering of Ministry of Education, Department of Biomedical Engineering, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, P. R. China.

Address correspondence and reprint requests to Mingxi Wan or Supin Wang, The Key Laboratory of Biomedical Information Engineering of Ministry of Education, Department of Biomedical Engineering, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, P. R. China. E-mail: mxwan@mail.xjtu.edu.cn or spwang@mail.xjtu.edu.cn

Journal of Voice, Vol. 27, No. 2, pp. 259.e7-259.e16
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2012.10.011

parameters (b/a) of phonation and silence are not linearity separable, so linear classification with a fixed single threshold could not absolutely distinguish one from the other, such as some closed-mouth phonemes mentioned by Wan et al.[17]

To realize an accurate voice initiation and termination, an artificial neural network (ANN) was introduced for switch identification and on/off control based on visual information of lips. ANN is a mathematical model widely applied in statistical pattern recognition.[18] The nonlinear nature of ANN will satisfy the mapping between lip features and voice on/off. Besides, the ANN has a strong robustness against noises. In this article, we implemented ANN method in a new video-controlled EL system (ANN-EL), which captured visual information of lips and controlled on/off of a wearable EL in real time. Furthermore, the performance of ANN method and ANN-EL were evaluated and compared with normal voice, button-EL, and video-EL.

METHODS

A schematic diagram of the experimental EL (ANN-EL) system is shown in Figure 1. The video signal of lips was captured and processed in real time to control on/off of EL. The procedure contained two main steps, which were lip-parameter extraction and on/off control.

Extraction of visual information of lips

There are two approaches widely used for extracting visual features, namely image-and model-based approaches. Considering the computational complexity and real-time implementation, model-based method was used to represent the lips by geometric parameters. The processes of parameters extraction were as follows: First, each frame of facial video image was preprocessed to decrease background noise and illumination. Second, the color image was filtered by a chromatic operator of lips and transformed to gray-scale image,[19] from which the lips was segmented by a threshold of gray-level histograms.[20] Finally, the lip outer contour was matched with an ellipse model and the shape parameters were extracted, namely semimajor (a) and semiminor axes (b).

On/off control with ANN

A two-layer feed-forward network was used in this article. It has been proved that with a sufficient number of hidden neurons, a multilayer perceptron neural network is capable of approximating an arbitrarily complex mapping within a finite support.[21] In the input layer, four inputs were normalized, namely semimajor axis (a/a_0), normalized semiminor axis (b/b_0), ratio of b to a (b/a), and normalized area of the ellipse (ab/a_0b_0). The parameters a_0 and b_0 represented the lip parameters of silence assigned during system initialization. Normalized parameters had advantages of distance and rotational invariances, so the influence of head movement could be reduced. For each neuron, the net function and the activation function were a weighted linear combination and a hyperbolic tangent activation function, respectively, which provided a nonlinear mapping between its input and output. The number of neurons was set as an empirical value of 20 in hidden layer. The network was trained using the scaled conjugate gradient back-propagation algorithm. The switch control depended on the two outputs, namely phonation and silence. The output of silence determined switch-off, and the other determined switch-on. Because the ANN algorithm was easy and fast, the real-time implementation was satisfied in our system.

The ANN-EL system

The ANN-EL system included three parts as shown in Figure 2. The first part was a microphone headset (Danyin DT-2699,

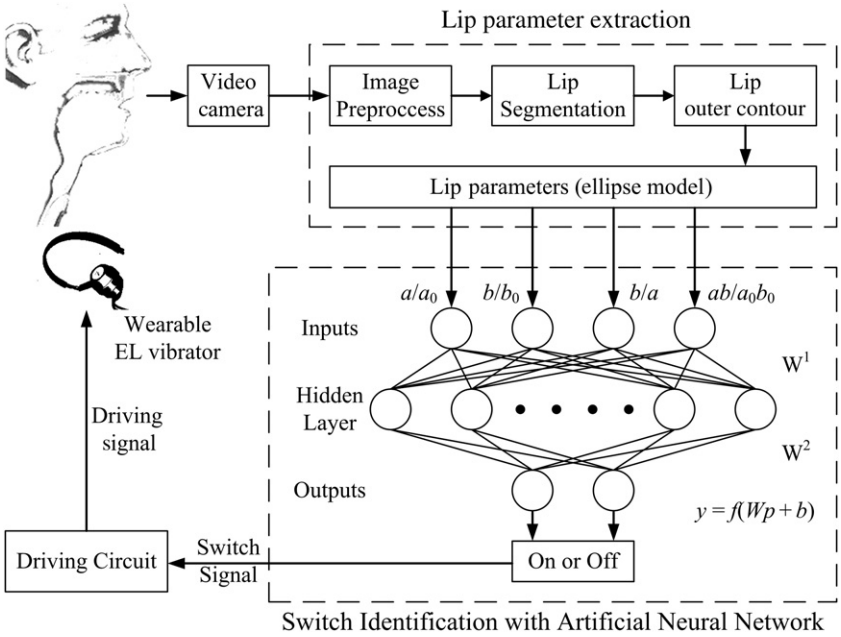

FIGURE 1. Schematic diagram of the ANN-EL system. The dash line boxes represent the extraction process of lip parameter and switch signal with artificial neural network.