Myo-To-Speech - Evolving Fuzzy-Neural Network Prediction of Speech Utterances from Myoelectric Signals

Mario Malcangi[1(✉)], Giovanni Felisati[2], Alberto Saibene[2], Enrico Alfonsi[3], Mauro Fresia[3], Roberto Maffioletti[1], and Hao Quan[1]

[1] Computer Science Department, Università degli Studi di Milano, Milan, Italy
{malcangi,maffioletti,quan}@di.unimi.it
[2] Department of Heath Science, Università degli Studi di Milano, Milan, Italy
{giovanni.felisati,alberto.saibene}@unimi.it
[3] IRCCS Mondino Foundation, Pavia, Italy
{enrico.alfonsi,mauro.fresia}@mondino.it

Abstract. Voice rehabilitation is needed after several diseases, when a subject's vocal ability is compromised by surgical interference or removal of phonation organs (e.g. the larynx), by neural degeneration or by neurological injury to the motor component of the motor-speech system in the phonation area of the brain (e.g. dysarthria in Parkinson disease). A novel approach to voice rehabilitation consists of predicting the phonetic control sequence of the voice-production apparatus (larynx, tongue, etc.) by drawing inferences on the basis of myoelectric (EMG) signals captured by a set of contact electrodes, applied to the neck area of a subject with important phonatory alteration (e.g. laryngectomised) and intact neural control. The inference paradigm is based on an EFuNN (Evolving Fuzzy Neural Network) that has been trained to use the sampled EMG signal to predict the phoneme that corresponds to the motor control of the sublingual muscle movements monitored at phonation time. A phoneme-to-speech synthesizer generates audio output corresponding to the utterance the subject has tried to enunciate.

Keywords: EFuNN · Evolving Fuzzy Neural Network · Voice dysarthria · Voice rehabilitation · Myoelectric signal

1 Introduction

Voice rehabilitation may be required in several conditions or post-surgical settings. The worst case scenario involving the phonation production organ is total laryngectomy. Voice rehabilitation for such patient is a challenging task [1] because of the total removal of this fundamental organ. On the other hand, when compared to other diseases that involve the voice production (e.g. neurological diseases) laryngectomised subjects maintain the integrity and the control of the remaining phonation organs, so a rehabilitation strategy could be implemented as non-surgical (esophageal speech or laryngophone) or surgical (esophageal puncture with voice prosthesis insertion).

E. Pimenidis and C. Jayne (Eds.): EANN 2018, CCIS 893, pp. 158–168, 2018.
https://doi.org/10.1007/978-3-319-98204-5_13

Laryngectomy drastically alters the speech production capabilities of human beings, since for speech production three main physiologic elements are necessary: the power source (lung air), the sound source (larynx) and the sound modifier (vocal tract). The only element that is active after laryngectomy is the sound modifier that is controlled by the brain.

Three main options are available to restore voice after laryngectomy: laryngophone speech (non-surgical, apparatus-based), esophageal speech (non-surgical, apparatus-free), and tracheoesophageal speech (surgical, hand-held device-based).

The non-surgical approach is interesting because it could be fully under the control of the subject without requiring hospitalization. In the esophageal speech production the air is swallowed into the esophagus and then released, inducing a vibration of the pharyngeal mucosa. This vibrations are modulated and articulated by the tongue movements and the control of the oral cavities.

This rehabilitation method is completely noninvasive, but it is difficult to learn (only 20% of laryngectomised patients succeed in this endeavor).

The laryngophone (or electrolarynx) is a non-surgical device-based approach that try to mimic electronically the larynx functionality of subjects who lost the larynx after the surgical removing.

The electrolarynx is a vibrating devices that is applied to the submandibular region and induces on the air in the oral cavities the vibration at a frequency mimicking that produced by the vibrating folds of the larynx. This fundamental vibration is modulated and articulated by the tongue and other mouth muscles producing an intelligible utterance.

The use of this machine requires a training phase and the use of the hands that holds the device in contact with the neck (it is almost impossible to talk to over the telephone). Other disadvantages concern the voce quality: the voice quality is metallic and unnatural. Furthermore, it is not applicable if the skin is not sound conductive. To minimize the disadvantages of the traditional electrolarynx device as voice prosthesis some investigations occurred in the past using electro-myoelectric activity to synchronize the device with the brain control of vocal tract at utterance-time [2].

The electrolarynx (Fig. 1) is the real demonstration that the best approach to voice rehabilitation in laryngectomised subjects consists in the reuse of the preserved voice control ability (vocal tract motion control). Following this idea we assumed that the myoelectric control of the phonation organs is driven by the phonetic information of the speech utterance. In the brain it exists a mapping between the language phonemes and the language words set, so when a word is to be uttered the corresponding phones sequence is predicted in terms of myoelectric control of voice articulation muscles (mainly the tongue).

If for each phoneme of an uttered word a corresponding myoelectric pattern sequence exists, then we could therefore theoretically predict such phoneme from the myoelectric pattern and use an articulatory phonetic speech synthesizer to electronically generate high-quality voice in laryngectomized subjects. This voice prosthesis would be hand-free operating, being fully brain-controlled.

Fig. 1. The Electrolarynx is a non-surgical device that mimics the larynx function in laryngectomised subjects by inducing air vibration in the vocal tract.

2 Physiology of the Phonation

Vocal production in mammals and humans originates from a complex mechanism. In a nutshell, the main mechanical interaction is between the air emitted by the lungs and then modulated through the main respiratory and accessory muscles (diaphragm, sternocleidomastoideus, intercostal muscles) and the action of the vocal cords and laryngeal structures on the expiratory flow [3, 4].

In humans the action of the intrinsic laryngeal muscles is rather complex and focuses on the rotation movements of the arytenoid cartilages, on the movements of antero-posterior displacement of the cricoaritenoid structures and on the variation of tension on the vocal cords, the latter resembling elastic bands with high vibratory capacity, able to develop vibrational frequencies that exceed even 100 Hz (Fig. 2). In particular, the abductor muscles of the vocal cords are the posterior cricoarytenoideus muscles; the tensor muscles of the vocal cords are tyroarynenoideus, Vocal and, mainly cricothyroid muscles; the adductor muscles of the vocal cords are the lateral cricoarytoenoideus and the interarytoenoideus muscles [5].

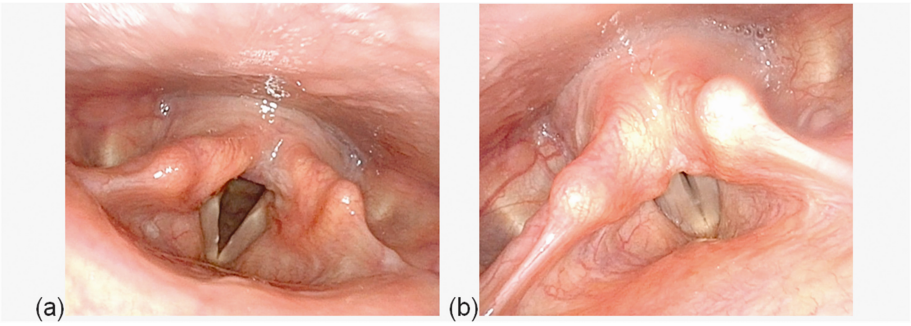

Fig. 2. Endoscopic view of an adult glottis, at rest (a) and during phonation (b), with medialization of the vocal folds in order to induce vibration and, consequently, sound formation.

The fundamental frequency and harmonics originated from the vocal folds vibration is then filtered inside the vocal tract (laryngeal cavity, pharynx, oral and nasal cavities) and modified by configurations and interactions of the articulators (e.g. tongue, lips and

palate) into producing speech sounds which are naturally linked into forming words and sentences [6].

3 Neurology of the Phonation

The emotional vocalization of the mammals seems to originate from the circuit that reaches the central ponto-bulbar "central pattern generators" [7, 8] from the cortex of the cingulate gyrus and the peri-aqueductal gray substance. This vocalization can be modified by environmental conditions but it cannot be learned, being instinctive and, probably, linked to mechanisms that are related to the survival of the species. These types of vocalization originate from the phylogenetically older structures of the cerebral cortex (paleocortex), present at the level of the limbic lobe and are connected to subcortical structures. Conversely, the ability to speak, that is language, is learned; it is generative rather than imitative, since the human species is able to formulate new sentences for communication.

The language is integrated with the auditory system and with the systems of voluntary motor control [9, 10]. Only the human species has a direct cortico-bulbar pathway that originates from the laryngeal motor cortex and reaches the ambiguous nucleus [11]. A clinical model particularly useful for understanding the different mechanisms between communicative language and emotional vocalization is represented by a neurological pathology called spasmodic dysphonia (DSP). In the DSP the neural systems of learning of the voluntary spoken language are involved, while those of emotional vocalization are not. The left cortical peri-silvial system, connected to the ability to develop the spoken language, according to the structural and functional magnetic resonance studies, consists of the cortical areas represented by the supragarginal tour, the arched fasciculus, the frontal opercular area M1 and the internal capsule [12, 13]. Laryngeal muscles are bilaterally controlled by both hemispheres [14] thus making the system vulnerable to unilateral abnormalities that interfere with bilateral control of laryngeal muscles. Diffusion MRI techniques have shown that fractional anisotropy is reduced in the knee region of the internal capsule in SDRs and there is an increase in diffusivity bilaterally at the level of the cortico-bulbar tract [13]. Even basal ganglia regions such as putamen, globus pallidus, substantia nigra, the posterior arm of the inner capsule, but also the locus coeruleus, show degenerative neuropathological changes in patients with DSP. These patients have also described inflammatory changes in the structures of the reticulofalic reticular substance of the word "central pattern generators", although this region represents the final common path of emotional and voluntary vocalizations and is strange that is involved in a pathology such as DSP in only the voluntary component is involved. According to some authors it is hypothesizable that the truncated dysentery anomalies are related to the fact that the structures affected here are a selective part of circuits used for the correct control and propagation of the verbal timing in the spoken language and are not involved in emotional vocalizations [15]. If, in the patient with DSP, attempts are made to satisfy precise requests for control of the single words in the spoken language, anomalous compensatory mechanisms may also develop in the systems of production of the word at the cortical level [15].

4 Capturing and Processing the Myoelectric Signal

The myoelectric signal was detected by two surface electrodes (Ambu Neuroline 715 silver-cloridate 10 mm x 6 mm) set placed over the skin of the suprahyoid subchin muscles. These muscles represent the main complex for the elevation of laryngeal-pharyngeal structures during swallowing and speech, the neck sublingual muscles (infrahyoid muscles) that depress the hyoid bone and larynx during swallowing and speech. The electrodes distance was 30 mm (15 mm lateral to the middleline of the neck of each electrode). A third electrode (reference - ground) was placed on the shoulder (clavicle). The three-electrode set (Fig. 3) detects in differential mode the electric potential that controls the muscle during utterance of each phoneme of the word. The electrodes are connected to a computer-based electromyograph (Viasys Healthcare's Medelec Synergy SYN5-C) to display and record the signals applying the following setup:

- Channel 1: connected to electrodes
- Preprocessing: rectification, band pass filtering – 100–2000 Hz.

The recorded myoelectric patterns have been exported from the memory electromyographer and processed by a Matlab application (software program) to window the pattern corresponding to the uttered phoneme.

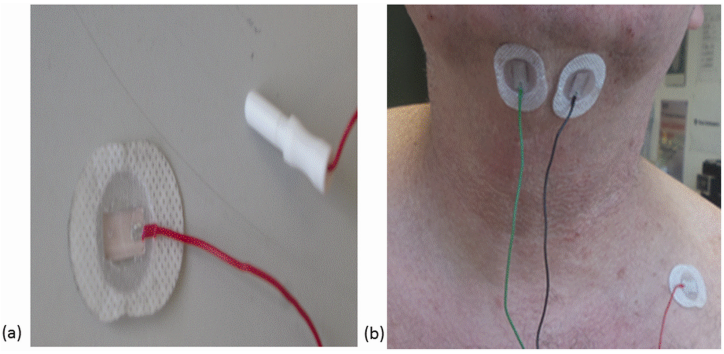

Fig. 3. Neurology surface electrode (Ambu Neuroline 715) (a) set placed over the skin of the suprahyoid subchin muscles and the reference electrode set placed over the clavicle (b).

5 Predicting from the Myoelectric Signal

The myoelectric sublingual muscle control signal embeds in its patterns the control sequence that moves the tongue to utter the word's phonemes. These patterns (Fig. 4) can be captured at speaking time and labeled with the corresponding phonemes to build up a labeled dataset to train a machine learning predicting paradigm.

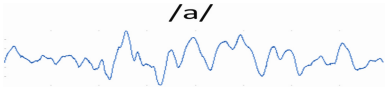

Fig. 4. Myoelectric signal captured synchronously at the utterance-time of the vocal /a/.

The training dataset consisted of the raw sampled data from the myoelectric signal and one label that classify the pattern as corresponding to a specific phoneme. Thousands of patterns need to be collected and classified to proceed to supervised training of the predicting paradigm (learning). After learning the paradigm will be able to predict from a myoelectric pattern the phoneme that the subject is to utter.

To accomplish this task, two main systems need to be deployed:

- The data acquisition of the myoelectric signal
- The predicting paradigm.

The data acquisition of the myoelectric signal is a challenging task due to the very low voltage physical nature of such signals, the noninvasive requirements of this medical but non clinical application and the high presence of artifacts in the captured signal.

The predicting paradigm is also challenging because the requirements cannot be accomplished by the hardcomputing digital signal processing algorithms.

Traditional softcomputing methods (Neural networks and Fuzzy Logic) demonstrated to be effective and powerful in solving nonlinear pattern matching issues but not adequate for on-line, life-long learning through adaptation in a changing environment. The new framework named Evolving Connectionist Systems (ECOS) [16], specifically the Evolving Fuzzy Neural Network (EFuNN), capable to build intelligent agents, is adequate to execute the phoneme's prediction from the myoelectric signal at silent phonation time.

6 The Evolving Fuzzy Neural Network (EFuNN) Paradigm

The EFuNN [17, 18] (Fig. 5) is a particular implementation of the ECOS [16] (Evolving COnnectionist System) a biologically inspired framework [16]. It is a softcomputing

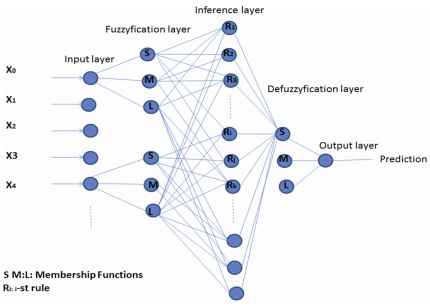

Fig. 5. EFuNN is a five layers Artificial Neural Network where each layer corresponds to a layer of a fuzzy logic engine.

paradigm that evolves through incremental, on-line learning, both supervised and unsupervised. EFuNN is order magnitude faster than multilayer perceptrons and fuzzy-neural networks.

EFuNNs are FuNNs that evolve according to the ECOS paradigm. FuNNs are Neural Networks that implements a set of fuzzy rules and the fuzzy inference engine in connectionist mode. FuNN is a feed-forward architecture-based with five layers of neurons and four layers of connections (Fig. 5). The first layer of neurons implements the input information layer. The second implements the membership layer that calculates the fuzzy membership degrees to which the input belongs to a fuzzy membership function. The third implements the fuzzy rules that encodes the associations between the input and the output data. The fourth calculates the degree to which the input data match the output membership functions. The fifth executes the defuzzification and calculates the crisp value for the output data. The FuNN is a combination of a Neural Network with a fuzzy engine The number of nodes and connections can change during the operation.

The peculiarity of EFuNN is its evolving capability and the one-pass learning. The nodes representing the membership function can be modified during the learning.

The rule nodes layer evolves through learning (supervised/unsupervised) that means all nodes are created/connected during learning. The nodes representing the membership functions can be modified at training-time. The same for the nodes representing the rules (input-output data association).

The evolving capability is incremental and adaptive. It is a bio-inspired way to make more effective the learning.

7 Dataset, Training and Test

To train and test the EFuNN's prediction capabilities related to the myoelectric patterns a dataset has been built. The dataset consists of several N-length raw sampled sequences of the myoelectric signal labeled by the corresponding phoneme code: $V_1\ V_2\ V_3\ V_4\ V_5\ V_6\ V_7\ V_8\ V_9\ V_{10}\ V_{11}\ V_{\ldots}\ V_{j\ \ldots}\ V_N\ L_n$

V_j: j-th amplitude of the j-th sample of the n-th myoelectric pattern
L_n: n-th label associated to the n-th sequence.

The N sequences composing the dataset are fed to the EFuNN settled in training mode. The EFuNN learns immediately from the data and it setups ready to be tested for prediction. A test dataset has been built in similar fashion of the training dataset. If the test is successful then the EFuNN is ready to run as phoneme predictor from myoelectric signal at silent speech production-time. If the test fails due to too many false predictions then the EFuNN is trained in evolving mode until the error rate is as low as required.

The first set of experiments to evaluate the EFuNN's phoneme prediction capabilities concerned the Italian language's vocal utterances (/a/ /e/ /i/ /o/ /u/ and the word /aiuola/).

The training dataset consists of myoelectric patterns captured and sampled synchronously with the utterance from an healthy subject.

The training consists only of the vocals, the test word is /aiuola/ that articulates almost all the vocals in a single word.

At training-time the EFuNN do not learn effectively. At test time it do not predict the right vocals sequence of the uttered word /aiuola/ under test (Fig. 6). After an evolving step (Fig. 7) the EFuNN learned to predict. At test-time it predicts without errors the right vocals phoneme sequence of the word /aiuola/: /a/ /i/ /u/ /o/ /a/ (Fig. 7).

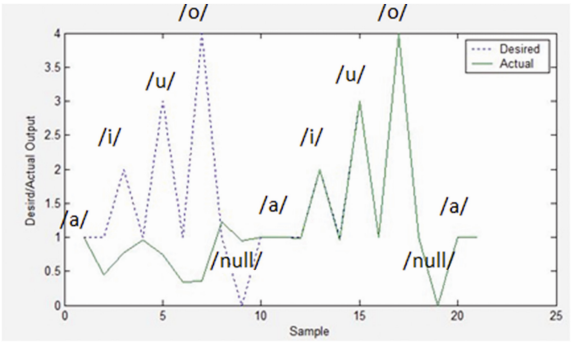

Fig. 6. Test /aiuola/ after one step training with single vocals /a/ /i/ /u/ /o/.

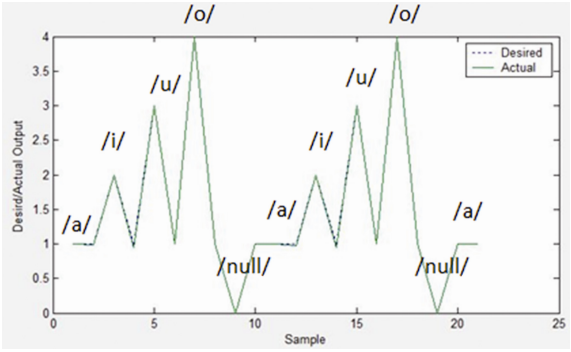

Fig. 7. Test /aiuola/ after one step evolving with single vocals /a/ /i/ /u/ /o/.

The modeling, training and test of the EFuNN have been executed with the simulation environment NeuCom developed at the Knowledge Engineering and Discovery Research Institute (KEDRI) Auckland – New Zeeland [19].

8 Myo-To-Speech System Framework

The Myo-to-Speech System Framework (Fig. 8) consists of three subsystems to transform the myoelectric signal from the sublingual muscle controlled by the phonation area of the subject's brain to the acoustical emission of the utterance.

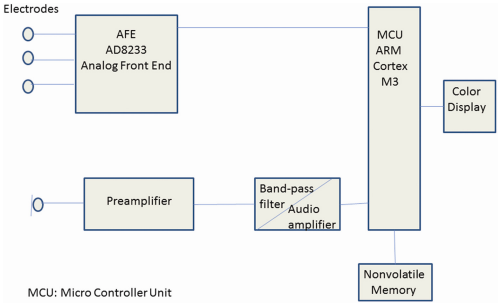

Fig. 8. Block chart of the prototyped myo-to-speech framework.

The first subsystem is the myoelectric signal acquisition AFE (Analog Front-End) consisting of the contact electrodes connected to an instrumentation amplifier and a set of filters to derive the bipolar electric signal from the differential electric signal captured by the contact electrodes.

The second subsystem is the Mixed-Signal MCU (Micro Controller Unit) that samples and collects in numeric format the myoelectric signal. This MCU executes also the inference paradigm (EFuNN) that predicts the phoneme from the myoelectric pattern.

The third subsystem is the phonemic speech synthesizer, that generates the electronic speech signal to be reproduced by a loudspeaker. The speech synthesizer is derived by the output of the predictor that encodes each predicted phoneme by a code embedded in the control part of the synthesizer.

9 Prototyping the Framework

To prototype the framework (Fig. 9) we used a set of fast prototyping COTS (Commercial Of The Shelf) boards. The most valuable part of the system is the analog front-end (AFE), that is available from Analog Devices in a CSP (Chip Scale Package) assembled on a prototyping COTS board (the AD8233 AFE). The AD8233 AFE is a fully integrated single lead electrocardiogram (ECG) analog front-end capable of 2–3 electrodes configuration with high signal gain (G = 100) with DC blocking capabilities and 80 dB (DC to 60 Hz) common rejection ratio. It integrates on a single-chip a 2-pole adjustable high-pass filter, one uncommitted operational amplifier and one 3-pole adjustable low-pass filter with adjustable gain.

A precision FET input unity-gain buffer (AD 8244) has been used to isolate source impedance from the of the signal chain.

Analog Devices AD8244's 2 pA maximum bias current, near zero current noise, and 10 TΩ input impedance introduce almost no error, even with source impedance well into the megaohms.

The AFE's input are connected to the three surface electrodes one reference and two differential. The AFE's output are connected to the bipolar input of the ADC (Analog to Digital Converter) of the MCU (MicroController Unit). The microcontroller unit is

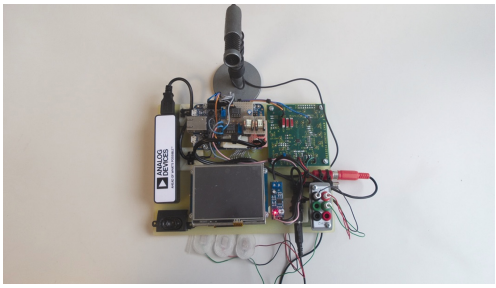

Fig. 9. The prototype integrates all the function blocks of the system framework using COTS

the NXP K64F an ARM Cortex M4F MCU running at 120 MHz, an ultra-low power processor integrating a rich set of mixed-signal peripheral and a huge amount of fast and non-volatile memory available for data acquisition and processing. The M4 processing core enables the computing intensive processing requirements without increasing of power consumption. The MCU is integrated on the evaluation board NXP Freedom K64 F.

A microphone (preamplified, band –pass filtered and amplified) is also connected to a separate input of the MCU's ADC for synchronous acquisition of the utterance with the myoelectric signal.

An SD non-volatile memory is used to storage the sampled myoelectric signal and the uttered signal. A TFT touch screen display controlled by an FTDI VM800C multimedia controller is used for signal acquisition monitoring and human-machine interface (HMI).

10 Conclusion and Future Developments

Compared to other similar approaches [20], our method is innovative as it refers to latest microelectronic technologies, to the most promising inference paradigm (evolving) and because its invasivity has been minimized. Methods like that proposed in [20] are largely invasive because multiple surface electrodes are applied to subject's face, and no wearable electronics has been developed to minimize the degree of invasiveness.

The first round of research and developments leads to the deploying of a prototype device that enables the Myoelectric-To-Speech (MyoToSpeech) synthesis for voice rehabilitation in laryngectomized subjects. The tests demonstrated that the predicting paradigm is effective but several issues need to be solved, concerning the automatic segmentation and labeling of the myoelectric signal, the porting of the electronics to a wearable (patch) size, the building of the datasets for different languages, to preserve and synthesize the subject's original voice.

Acknowledgments. A special acknowledgment is due to Prof. Nikola Kasabov, Auckland University of Technology, Director KEDRI – Knowledge Engineering and Discovery Research Institute, for his invaluable suggestions on how to get the most from the EFuNN's evolving capabilities.

Acknowledgment is also due to Jan Hein Broeders (Analog Devices' healthcare business-development manager for EMEA) for his precious support and expertise in hardware prototyping, especially for the analog front-end (AFE) subsystem.

References

1. Balasubramanian, T.: Voice rehabilitation following total laryngectomy. Otoryngology Online J. **5**(1), 5 (2015)
2. Goldstain, E.A., Heaton, J.T., Kobler, B., Stanley, G.B., Hillman, R.E.: Design and implementation of a hand-free electrolarynx device controlled by neck strap muscle electromyographic activity. IEEE Trans. Biomed. Eng. **51**, 325–332 (2004)
3. Titze, I.R.: The physics of small-amplitude oscillation of the vocal folds. J. Acoust. Soc. Am. **83**, 1536–1552 (1988). https://doi.org/10.1121/1.395910. PMID 3372869
4. Lucero, J.C.: The minimum lung pressure to sustain vocal fold oscillation. J. Acoust. Soc. Am. **98**, 779–784 (1995)
5. Lucero, J.C.: Optimal glottal configuration for ease of phonation. J. Voice **12**, 151–158 (1998)
6. Mor, N., Simonyan, K., Blitzer, A.: Central voice production and pathophysiology of spasmodic dysphonia. Laryngoscope **128**(1), 177–183 (2018). Epub 23 May 2017, Review (2018)
7. Jürgens, U.: Neural pathways underlying vocal control. Neurosci. Biobehav. Rev. **26**, 235–258 (2002)
8. Jürgens, U.: A study of the central control of vocalization using the squirrel monkey. Med. Eng. Phys. **24**, 473–477 (2002)
9. Vihma, Mn., de Boysson-Bardies, B.: The nature and origins of ambient language influence on infant vocal production and early words. Phonetica **51**, 159–169 (1994)
10. Mac Neilage, P.F.: The frame/content theory of evolution of speech production. Behav. Brain Sci. **21**, 499–511 (1998)
11. Kuypers, H.G.: Cortico-bulbar connexions to the pons and lower brainstem in man: an anatomical study. Brain **81**, 364–388 (1958)
12. Haslinger, B., Erhard, P., Dresel, C., Castrop, F., Roettinger, M., Ceballos-Baumann, A.O.: Silent event-related, fMRI reveals reduced sensorimotor activation in laryngeal dystonia. Neurology **65**, 1562–1569 (2005)
13. Simonyan, K., Ludlow, C.L.: Abnormal activation of the primary somatosensory cortex in spasmodic dysphonia: an fMRI study. Cereb. Cortex **20**, 2749–2759 (2010)
14. Rödel, R.M., et al.: Human cortical motor representation of the larynx as assessed by transcranial magnetic stimulation (TMS). Laryngoscope **114**, 918–922 (2004)
15. Ludlow, C.L.: Spasmodic dysphonia: a laryngeal control disorder specific to speech. J. Neurosci. **31**(3), 793–797 (2011)
16. Kasabov, N.: Evolving Connectionist Systems: The Knowledge Engineering Approach. Springer, Heidelberg (2007). https://doi.org/10.1007/978-1-84628-347-5
17. Kasabov, N.: EFuNN, IEEE Tr SMC (2001)
18. Kasabov, N.: Evolving fuzzy neural networks – algorithms, applications and biological motivation. In: Yamakawa, T., Matsumoto, G. (eds.) Methodologies for the Conception, Design and Application of the Soft Computing, pp. 271–274. World Computing (1998)
19. http://www.kedri.aut.ac.nz/areas-of-expertise/data-mining-and-decision-support-systems/neucom
20. Zahner, M., Janke, M., Wand, M., Schultz, T.: Conversion from facial myoelectric signals to speech: a unit selection approach. In: Interspeech, pp. 1184–1188 (2014)