Speech Driven by Artificial Larynx: Potential Advancement Using Synthetic Pitch Contours

Hua-Li Jian(✉)

Institute of Information Technology, Faculty of Technology, Art and Design
Oslo and Akershus University College of Applied Sciences, Oslo, Norway
Hua-Li.Jian@hioa.no

Abstract. Despite a long history of development, the speech qualities achieved with artificial larynx devices are limited. This paper explores recent advances in prosodic speech processing and technology and assesses their potentials in improving the quality of speech with an artificial larynx – in particular, tone and intonation through pitch variation. Three approaches are discussed: manual pitch control, automatic pitch control and re-synthesized speech.

Keywords: Artificial larynx · Fundamental frequency · Assistive technology

1 Introduction

Some individuals lose their ability to produce vowels after having their larynx surgically removed, for instance, after cancer in the throat. The larynx, or voice-box, produces the sound in the throat that drives speech. Individuals who have had their larynx removed can learn to produce esophageal speech, that is, the oscillation of the esophagus. Esophageal speech requires training and is strenuous and has a limited volume. Another approach is to surgically implant voice prosthetics. This paper focuses on the non-surgical approaches based on the electrolarynx. The advantages of eletrolarynxes are that very long sentences can be produced. There is no need for surgical procedures, and the device can be operated with virtually no skill and no maintenance.

There are various types of artificial larynxes or electrolarynxes. Most artificial larynxes are handheld devices held towards the throat. The device generates a vibration that is directed towards the throat that the speaker can use as basis for generating vowels in addition to consonants, in particular, plosives which are generated without the larynx.

Artificial larynxes have a push-to-talk button, and some designs have also a pitch control. An issue with artificial larynxes is the lack of naturalness and research has thus gone into assessing their naturalness [1, 2]. The interaction of a speaker and the chosen artificial larynx may also affect the intelligibility of the speech realized. Factors including gender, physiological states, and user proficiency may all impact such realization. Formalized tests concerning speech intelligibility and acceptability were thus advised for individual users prior to settling the most suited artificial device [1].

Artificial larynx speech can be stigmatizing for its users due to the highly noticeable monotone speech. Such speech has been used as characteristics in popular culture such

M. Antona and C. Stephanidis (Eds.): UAHCI 2015, Part III, LNCS 9177, pp. 312–321, 2015.
DOI: 10.1007/978-3-319-20684-4_30

as Ned Gerblansky in the South Park TV-series or Charlie in the Mad Max movie. An individual with an artificial larynx may not be taken seriously on the phone if the talker at the other end does not know that the speaker uses an artificial larynx. Moreover, as the speech is harder for untrained listeners to understand, miscommunication can occur, especially in a noisy environment such as a public space. By striving towards more natural sounding speech, it is likely that both the stigma can be reduced and the communication with others improved.

Although some development has been made over the course of nearly 120 years, the amount of research into artificial larynxes is limited. One possible explanation for the lack of attention could be that the proportion of individuals dependent on artificial larynxes are relatively limited with perhaps less than 100 per million people. The objective of this paper is to explore the possible use of recent technological developments and off-the-shelf third party technology intended for other purposes to improve the quality of speech by individuals without a natural voice box.

2 Background

Surgery of the larynx can lead to partial or full disability to produce vowels [3]. This study focuses on individuals who are reliant on artificial larynxes to produce speech. The waveform produced by the human larynx is complex and some research has gone into understanding the underlying mechanisms leading to the rich timbre produced by the larynx [4]. Some researchers have also attempted to improve the spectrum produced by the electronic larynx by the means of piezoeletronic ceramics as the source of the vibrations [5].

The artificial larynxes are also known to produce harsh background noises that reduce the speech quality, and measures to reduce the noise have been made using adaptive noise cancellation. The processes also helped preserve the voice's acoustic characteristics and hence speech acceptability was improved [6].

Recent technology allows researchers to focus even more on the naturalness of the actual vibrations through accurate observation of the larynx using high speed video [7]. With such objectively quantified information, further rehabilitation of the substitute voice may be achieved.

Without a larynx, individuals are able to whisper without the aid of an artificial larynx. However, whispering is usually too weak in volume to be practical in everyday conversation. To overcome these problems, researchers have also attempted to capture whispered speech and re-synthesize normal speech externally [8]. However, this approach is sensitive to background noise in the environment.

Alternative means of controlling the artificial larynx have also been used, such as employing the myoelectric signals that can be measured around the neck [9, 10] to control both push-to-talk and pitch at low, medium and high frequencies. Experiments with wireless connections between the neck sensors and the vibrator have also been explored [11].

Pitch has been identified as a key characteristic of speech. In one approach, air pressure through breathing was used to control the pitch of an artificial larynx [12].

This setup required training and users' training needs might vary. The intonation of a short sentence thus produced was reported to be similar to that of a normal subject.

3 Pitch Control by Manual Adjustments

The idea of manual pitch control for artificial larynxes is not new. Initially, some artificial larynx devices were designed with a pitch control. Recent studies have explored hands-free approaches to pitch control using breathing pressure [12] and myoelectric signals [9].

The monotonic voice of synthetic speech for disabled users has troubled researchers. One strategy proposed allows typically non-technical users to transcribe prosodic features of speech for artistic performances off-line in configuration files [13]. However, this is more suitable for users relying on synthetic speech and not users relying on artificial larynxes.

With current mobile technology enabling real world gestures, a specialized handheld mobile device or a general one such as a smartphone could be used to express gestures that again would control the pitch of the device. For instance, a rising pitch could be achieved by lifting the device, while a falling pitch could be achieved by lowering the device (see Fig. 1).

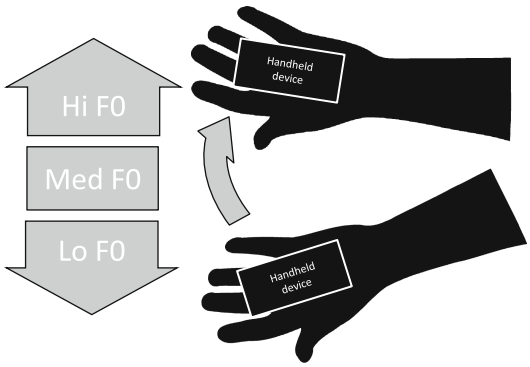

Fig. 1. Lowering and lifting the device to control pitch (f0) height

Most smartphones are equipped with accelerometers that are able to accurately detect such gestures [14]. Wireless communication, such as Bluetooth, could be used to send the pitch information from the handheld device to the artificial larynx. Such gesticulation control would not be subject to the gesture segmentation problem that is present in other application areas where gestures are used [15].

This approach would be relatively simple to implement from a technical standpoint. It would give the speaker complete control over the pitch, although some hand-pitch coordination training might be needed. The approach would be particularly effective in performance settings, e.g., during public speaking where gesticulations and exaggerated prosody often are used to emphasize the message. On the other hand, this approach

may not be as effective, and even disturbing, during more calm settings such as when conversing in a small group. It may also be strenuous to physically articulate pitch with the hands for long periods of time. Automatic pitch enhancements are thus explored in the subsequent sections.

4 Pitch Control by Intelligent Systems

An alternative to manual pitch control is automatic pitch control. One may hypothesize that imposing randomly selected natural pitch curves to the speech would improve its naturalness by breaking the monotony. By employing an even more intelligent strategy, that is, choosing a more fitting pitch curve than random is likely to yield better results. We can draw inspiration from text-to-speech systems.

Speech quality is found to be most sensitive to pitch in text-to-speech systems [16, 17]. The addition of pitch accents to synthetic speech has long been a research topic with many innovative approaches. For instance, Hidden Markov models are applied to estimate the prosodic features of synthetic speech [18]. Researchers have also examined the detailed assignment of pitch marks at waveform level [19].

Simple rules have also been used to set the pitch accent of synthetic speech. Hirschberg [20] proposed the following rules: cue phrases (e.g., *now, we,* and *by the way*) are key accented, closed classed words are de-accented, words with their root in local or global focus are de-accented, compound stress assignments suggesting de-accenting are de-accented, and all other cases are accented. Algorithms may also be created to infer topic structure from paragraphing, punctuation, and lexical cues [21]. For achieving speech naturalness and successful listeners' interpretation, accent assignment denoting which words to emphasize or de-emphasize intonationally is important. Recent experiments on recorded read speech and elicited speech have demonstrated considerable success (over 80 % correctness) in modeling speakers' accenting strategies by merely using automated text analysis [22].

Most of the work on prosody for text-to-speech systems is based on pitch measurements for various transcribed speech corpora, that is, pitch extracted from authentic speech. The pitch contours can be associated with single vowels, words or phrases, and sometimes combined with sentence templates. It is likely that a prosodic module from a text-to-speech system could be adopted to the automatic pitch control of artificial larynxes with relatively moderate effort.

4.1 Acquiring Speech Information

A key difference between a text-to-speech system versus an artificial larynx is the lack of information available. With a text-to-speech system, the text to be uttered is known a priori. With an artificial larynx. there is no basic information available. However, the following information could be solicited.

Segment durations: the user of the artificial larynx controls the device with the on/off switch. The state of the on/off switch provides useful information in terms of when speech is uttered and when it is not. Moreover, the timing of the speech

segments, that is, the speech duration, and the durations of the pauses, could provide useful cues.

Audio: the resulting speech produced with the artificial larynx could be recorded in real time and subjected to speech recognition technology. This could be achieved by attaching a microphone to the artificial larynx device. For the system to work in real time, the recognition would have to be at the level of phonetic units. The feasibility and accuracy of acquiring phonetic information from speech driven by an artificial larynx would need to be investigated. However, there is potential to specially train a speech recognition engine for such speech.

Neck muscle signals: by attaching sensors at the neck, valuable information about the throat muscle activity could be measured to help classify the uttered sound. Such signals could be used together with audio signals to improve the recognition rates.

Image data: to further help the real time recognition of the uttered signals, visual cues could be acquired using a video camera. Research into video analysis has successfully managed to lip-read utterances simply from visual cues in color videos [23–25]. By combining several channels such as audio, video and muscle information, a more accurate phonetic classification may be achieved. One challenge is where and how to fit the camera to obtain high recognition rates while ensuring sufficient pervasiveness in the setup.

4.2 Speech Prediction

As the pitch has to be adjusted in real time, partially uttered speech needs to be used to predict the intended utterances. For this purpose, text prediction algorithms [26] such as trie structures can be employed. Text prediction algorithms in the simple form can be composed using di-grams, where pairs of phonetic elements comprise each di-gram key and the di-gram entry is assigned a pitch contour. Table 1 shows an extract of such a di-gram. Next, imagine that the first syllable of an utterance is A and the second is D, with the corresponding pitch contour being LEVEL. The trie approach involves a linguistic model using a dictionary with all forms of the words organized into a tree-like structure [27].

Table 1. Example syllable di-gram extract with pitch contours

1st syllable	2nd syllable	Pitch contour
A	B	FALL
A	D	LEVEL
A	E	RISE
…	…	…

For the approach to work, the data structure needs to be based on phonetic elements rather than spelling. One benefit of pitch prediction over spelling prediction is that the number of unique pitch patterns is smaller than the set of possible spellings. Examples include using the simple SOUNDEX or more sophisticated metaphone strategy [28];

the latter is commonly employed for phonetic matching in spelling correction applications. Instead of mapping the partial utterance with a particular word, it is associated with a given frequency contour. Interpolation of pitch contours can be used to make a smooth switch from an incorrectly predicted contour to the intended contour in erroneous cases. As predictions are based on partial utterances, the prediction accuracy will be lower in the beginning of an utterance compared to when the utterance is complete. If there is a tie between several pitch contours, the most probable contour can be selected.

Table 2 shows an example of pitch prediction where the first syllable of the utterance is Y and thus assigned a mid-level pitch contour with a low confidence of 10 %. The second syllable is E giving the prefix YE, which means that the pitch contour prediction is altered to a rising contour with 25 % confidence.

Table 2. Example of observation window, prediction and confidence

Observation	Prediction	Confidence
Y*********	mid-level	10 %
Ye********	Rise	25 %
Yes*******	Rise	50 %
Yes *******	Rise + pause	100 %

Next, the third syllable is S giving the prefix YES, which also is assigned to a rising pitch contour with a confidence of 50 %. Finally, a pause is detected and the uttered word is detected as YES with a rising pitch contour.

A simple selection scheme was also proposed to reduce mismatch of pitch and thus increase pitch prediction rate [17]. By means of annotations employing linguistic foot structure, local pitch contours of syllables could be predicted more accurately.

4.3 Speech Correction

Misrecognition for any spoken system is to be avoided or corrected. Strategies for correcting, rejecting, or changing misrecognized hypotheses have been proposed [29–31]. Prosodic features such as F0 perturbation, duration, and loudness were shown to significantly characterize failed recognition runs in terms of word-accuracy and conception-accuracy [32]. Machine learning experiments also indicated that use of prosodic differences may greatly improve prediction of misrecognition in terms of word-accuracy and obtain even greater predication rate when combining prosodic features with other automatically available features of speech recognition systems.

Variation of speaking rate may have a negative impact for automatic recognition systems [33]. Possibly longer utterances, varied or irregular pausing, and slow articulation combined with disfluency may all cause recognition errors [32]. Understandably, such chances of error may be even higher for individuals using artificial larynxes.

It may, however, be possible to make users aware of recognition errors [34] and also correct them [35]. Efforts into examining prosodic variations have been made to

account for why some voices are more poorly recognized than others [36, 37]. Failure identification and reaction strategies in speech recognition systems may be enhanced by integrating prosodic-related information [32].

Correcting misrecognitions by users have also been predicted and analyzed. User corrections were more poorly recognized than non-corrections, but they were not more frequently rejected by the recognition systems. Corrections paraphrasing the original information were found to be less recognized than those omitting it [38]. However, user corrections were found to be better identified by means of a combined feature set of prosody and specific system-derived features. Future techniques to improve correction prediction and to further execute modifications for automatic correction identification would also add to positive development of speech recognition for artificial-larynx-driven speech.

5 Total Synthetic Speech

One could imagine going one step further by synthesizing the speech in real time based on the successfully recognized utterances. In this case the artificial larynx would be replaced by an artificial voice altogether using synthetic speech. In this way it may be possible to somehow restore the original voice of someone who has undergone surgery on the larynx. Some research has also gone into synthesizing speech with arbitrary voices [39].

However, such an approach raises new issues such as where the sound should come from. An advantage of the artificial larynx is that the sound still originates from the throat of the speaker. With synthetic speech, it is important that the speaker is as close as possible to the speaker's mouth to give the impression that the sound actually originates from the speaker. Otherwise, if the speech comes from a different location, the listeners may get confused in conversational settings. Moreover, this scheme would require highly accurate speech recognition. The impact of misrecognition consequently resulting in erroneous speech synthesis is more severe than an incorrect pitch contour, which in the worst case will only sound odd.

One major issue with synthetic speech systems is potential lags caused by processing delays as the detection and synthesis involved are in essence complex operations.

6 Summary and Future Work

This paper has explored the problem of lacking prosody in speech produced with artificial larynx devices. Three approaches to improving the expression pitch for such speech using recent technological advances and off-the-shelf hardware are discussed, namely, the simplest strategy of manual control of pitch through gestures via a handheld device, the automatic control of pitch via speech recognition, and finally the most challenging idea of total real-time synthesis of speech based on real-time speech recognition.

The simple approach such as the manual control would probably be associated with an unperceivable delay. With efficient algorithms and high performance hardware, it may be possible to reduce processing lags to a minimum. Future work will focus on (a) exploring the perception effects of altering the pitch and (b) developing a robust pitch-contour prediction algorithm.

References

1. Stalker, J.L., Hawk, A.M., Smaldino, J.J.: The intelligibility and acceptability of speech produced by five different electronic artificial larynx devices. J. Commun. Disord. **1**(5), 299–301 (1982)
2. Pindzola, R.H., Moffet, B.: Comparison of ratings of four artificial larynxes. J. Commun. Disord. **21**, 459–467 (1988)
3. Modrzejewski, M., Olszewski, E., Wszol, W., Rerona, E., Strek, P.: Acoustic assessment of voice signal deformation after partial surgery of the larynx. Auris Nasus Larynx **26**, 183–190 (1999)
4. Alipour, F., Scherer, R.C., Finnegan, E.: Measures of spectral slope using an excised larynx model. J. Voice **26**(4), 403–411 (2012)
5. Ooe, K., Fukuda, T., Arai, F.: A new type of artificial larynx using a PZT ceramics vibrator as a sound source. IEEE/ASME Trans. Mechantronics **5**(2), 221–225 (2000)
6. Niu, H.J., Won, M.X. Waq, S.P.: Enhancement of electronic artificial larynx speech by denoising. In: IEEE International Conference on Neural Networks & Signal Processing, pp. 908–911. IEEE Press (2003)
7. Schwarz, R., Huttner, B., Dollinger, M., Luegmair, G., Eysholdt, U., Schuster, M., Lohscheller, J., Gurlek, E.: Substitute voice production: quantification of PE segment vibrations using a biomechanical model. IEEE Trans. Biomed. Eng. **58**(10), 2767–2776 (2011)
8. Sharifzadeh, H.R., McLoughlin, I.V., Ahmadi, F.: Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec. IEEE Trans. Biomed. Eng. **57**(10), 2448–2458 (2010)
9. Ooe, K.: Development of controllable artificial larynx by neck myoelectric signal. Procedia Eng. **47**, 869–872 (2012)
10. Stepp, C.A., Heaton, J.T., Rolland, R.G., Hillman, R.E.: Neck and face surface electromyography for prosthetic voice control after total laryngectomy. IEEE Trans. Neural Syst. Rehabil. Eng. **17**(2), 146–155 (2009)
11. Heaton, J.T., Robertson, M., Griffin, C.: Development of a wireless electromyographically controlled electrolarynx voice prosthesis. In: 33rd Annual International Conference of the IEEE EMBS, pp. 5352–5355. IEEE Press (2011)
12. Uemi, N., Ifukube, T., Tamashi, T., Matsushima, J.: Design of a new electrolarynx having a pitch control function. In: IEEE lnternational Workshop on Robot and Human Communication, pp. 198–203. IEEE Press (1994)
13. Blankinship, E., Beckwith, R.: Tools for expressive text-to-speech markup. In: Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology, pp. 159–160. ACM press (2001)
14. Győrbíró, N., Fábián, A., Hományi, G.: An activity recognition system for mobile phones. Mobile Netw. Appl. **14**(1), 82–91 (2009)

15. Carrino, F., Ridi, A., Ingold, R., Abou Khaled, O., Mugellini, E.: Gesture vs. gesticulation: a test protocol. In: Kurosu, M. (ed.) HCII/HCI 2013, Part IV. LNCS, vol. 8007, pp. 157–166. Springer, Heidelberg (2013)
16. Plumpe, M., Meredith, S.: Which is more important in a concatenative text to speech system - pitch, duration or spectral discontinuity? In: Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan, Australia (1998)
17. Klabbers, E., van Santen, J.P.H.: Control and prediction of the impact of pitch modification on synthetic speech quality. In: Eurospeech 2003 (2003)
18. Gu, H.Y., Yang, C.C.: An HMM based pitch-contour generation method for mandarin speech synthesis. J. Inf. Sci. Eng. **27**, 1561–1580 (2011)
19. Chen, J.H., Kao, Y.A.: Pitch marking based on an adaptable filter and a peak-valley estimation method. Comput. Linguist. Chin. Lang. Process. **6**(2), 1–12 (2012)
20. Hirschberg, J.: Accent and discourse context: assigning pitch accent in synthetic speech. In: AAAI 1990 Proceedings (1990)
21. Hirschberg, J., Litman, D.: Disambiguating cue phrases in text and speech. In: Proceedings of COLING 1990, Helsinki, August (1990)
22. Hirschberg, J.: Pitch accent in context predicting intonational prominence from text. Artif. Intell. **63**(1), 305–340 (1993)
23. Chiou, G.I., Hwang, J.N.: Lipreading from color video. IEEE Trans. Image Process. **6**(8), 1192–1195 (1997)
24. Zhou, Z.H., Zhao, G.Y., Pietikainen, M.: Towards a practical lip-reading system. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 137–144 (2011)
25. Li, M., Cheung, Y.M.: A novel motion based lip feature extraction for lip-reading. In: International Conference on Computational Intelligence and Security, CIS 2008, vol. 1, pp. 361–365 (2008)
26. Garay-Vitoria, N., Abascal, J.: Text prediction systems: a survey. Univers. Access. Inf. Soc. **4**(3), 188–203 (2006)
27. Fredkin, E.: Trie Memory. Commun. ACM **3**(9), 490–499 (1960)
28. Philips, L.: Hanging on the metaphone. Comput. Lang. **7**(12), 38–43 (1990)
29. Litman, D., Walker, M., Kearns, M.: Automatic detection of poor speech recognition at the dialogue level. In: Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics, ACL 1999, College Park, pp. 309–316 (1999)
30. Litman, D., Pan, S.: Empirically evaluating an adaptable spoken dialogue system. In: Proceedings of the 7th International Conference on User Modeling (UM), Banff, pp. 55–64 (1999)
31. Walker, M., Kamm, C., Litman, D.: Towards developing general models of usability with PARADISE. Nat. Lang. Eng. Special Issue on Best Practice Spoken Language Dialogue System Engineering **6**, 363–377 (2000)
32. Hirschberg, J., Litman, D., Swerts, M.: Prosodic and other cues to speech recognition failures. Speech Commun. **43**(1), 155–175 (2004)
33. Ostendorf, M., Byrne, B., Bacchiani, M., Finke, M., Gunawardana, A., Ross, K., Roweis, S., Shriberg, E., Talkin, D.,Waibel, A., Wheatley, B., Zeppenfeld, T.: Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. In: Report on 1996 CLSP/JHU Workshop on Innovative Techniques for Large Vocabulary Continuous Speech Recognition (1997)
34. Litman, D., Hirschberg, J., Swerts, M.: Predicting user reactions to system error. In: Proceedings of the ACL-2001, Toulouse, pp. 329–369 (2001)
35. Hirschberg, J., Litman, D., Swerts, M.: Identifying user corrections automatically in spoken dialogue systems. In: Procedings of the NAACL 2001, Pittsburgh, pp. 208–215 (2001)

36. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D.: Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In: Proceedings of the International Conference on Spoken Language Processing-98, Sydney, pp. 608–611 (1998)
37. Hirschberg, J., Litman, D., Swerts, M.: Prosodic cues to recognition errors. In: Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU 1999), Keystone, pp. 349–352 (1999)
38. Litman, D., Hirschberg, J., Swerts, M.: Characterizing and predicting corrections in spoken dialogue systems. Comput. Linguist. **32**(3), 417–438 (2006)
39. Tamura, M., Masuko, T., Tokuda, K., Kobayashi, T.: Text-to-speech synthesis with arbitrary speaker's voice from average voice. In: Proceedings of Eurospeech 2001, pp. 345–348 (2001)