Matlab Tool for Assistance in Learning English Words Pronunciation

Jan Malucha

Brno University of Technology, Czech Republic

E-mail: xmaluc00@vutbr.cz

Abstract—This paper presents a MATLAB module to assist students in learning the correct pronunciation of individual English words. The measurement of pronunciation correctness is based on evaluation of prosodic speech parameters using three methods. A brief introduction to these prosodic parameters is given, along with a description of the developed tool. Program testing was performed with a native and non-native speaker.

Keywords—speech signal, pronunciation, prosody

1. INTRODUCTION

Verbal communication is one of the key elements of social life. Anglocentrism has been unrivaled in global communication at all levels since the 1990s. The English language has taken on the role of the new lingua franca, whose knowledge determines access to information and opportunities; this trend is likely to continue. However, research by the Czech Statistical Office shows that only 42% of Czech citizens aged 18 to 69 speak English, with most of this percentage having only a basic or minimal knowledge of English [1]. Since most Czech users of English as well as other foreign languages very rarely get an opportunity to converse with a foreign speaker, it can be assumed that the ability to pronounce correctly (and thus communicate verbally) in a foreign language is even lower in the Czech Republic than the reading skills. According to [2], a very narrow space is also devoted to the teaching of pronunciation in education. However, given the growing trend of international connection and the growing importance of English, it makes great sense to focus on teaching pronunciation, which also encourages efforts to develop software tools to support and streamline this teaching. A number of such tools have been developed over the years, with attention given to them in [3] or [4], for example.

2. PROSODY

From the point of view of phonetics, the basic means of speech communication is a continuous sound stream. From the acoustic point of view, it is a vibration of the transmission medium excited by the speech system, but for the analysis of speech signal, it is understood as a chain of noise and tone acoustic components forming individual sounds - vowels. Thus, the vowel is an elementary segment of speech and the suprasegmental segment is a syllable [5]. We then understand prosody as a set of properties of speech signal segments related to the speech modification, or speech flow modulation. This modulation is realized through the so-called modulation factors - accent and rhythm of speech are given by the factor of strength or **energy**, tempo and rhythm are given by the factor of **duration**, tone and intonation are given by the factor of **pitch**. Since there are some typical attributes related to these factors in a language, it is possible to analyze the correctness of the pronunciation by determining these parameters during the speech signal. The meaning of prosody is described in [6]. Statistics of voiced speech can be seen in [7].

3. AIMS

The aim of this paper is to present our MATLAB tool developed for analysis of the input speech signal of a *student's* word along with its processing and evaluation with respect to the correctness of pronunciation. The prosodic criteria on which this analysis is based are **energy**, **duration**, and **pitch**, from which information about speech accent, melody, and length of word segments is derived. Accuracy assessment is based on a comparison with the *lecturer's* reference word. The output for practical usability is simple graphical and verbal feedback to the pronunciation, referring to the accented phonemes, word melody and segment lengths.

4. TOOL DESCRIPTION

The program is based on the method of short-time analysis. At first, the input signal is divided into short-time segments of 20 ms and the processing then takes place individually in these segments, thus achieving the stationarity of the processed speech data. The overall structure is outlined in Fig. 1. Because two signals (*student* and *lecturer*) are compared and the difference in duration of both signals is assumed due to faster or slower speech, both signals are normalized to the percentage axis. The tool itself consists of three main algorithms supported by a number of auxiliary algorithms. It was implemented in MATLAB.

The method used for **energy** analysis is STE (short time energy). This procedure determines the energy of the short-time segment by calculating the power and the sum [8]. It is built into a complex algorithm, the input of which is a speech signal and the output is both a graph of energy along the signal and a table comparing the energies of selected sections of the *student's* and *lecturer's* signal. Thanks to this, it is possible to provide information about the difference of the accent, i.e. accented syllables, between the *student's* and *lecturer's* signals. The **duration** analysis also uses STE methods, this time built into a second algorithm, which identifies voiced, unvoiced and silent sections of the signal by comparing the course of energy over time with experimentally selected thresholds. These parts can be understood as a precursor to phonemes; the algorithm then determines the positions of the voicing transitions in time, compares these transitions for the *student* and *lecturer* signal in the table and gives information about the duration of individual sections. The **tone** along the time axis, i.e. melody, is determined in the third algorithm using the time domain method. Fourier transform including modifications [9] and autocorrelation methods [10] have traditionally been used to estimate frequency in speech. The computational procedure AMDF [11] is used here, which is based on the autocorrelation function. This method is built into an algorithm, the output of which is a graph of the normalized pitch course centered around its mean value. It can clearly show the positions of raises and drops of the voice along the timeline – the up-crossing of the mean value is considered a voice raise, down-crossing indicates the opposite. These increases and decreases are then located, saved and compared in a table for both the *student* and *lecturer* signal, and a brief verbal feedback about the melody is provided to the user.

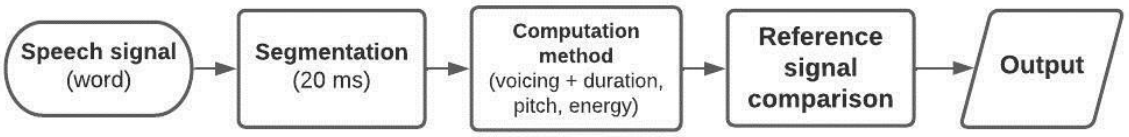

Figure 1: Program flowchart

5. PRACTICAL APPLICATIONS AND RESULTS

The practical significance of the tool lies in the support of independent pronunciation learning of English words without the presence of a teacher. The English language is relatively sensitive to correct intelligibility, and incorrect articulation or duration of the sound can often cause a complete change in the meaning of the word (e.g. the difference between the words "hid" and "heed". Intonation, as a combination of melody and accent, is also one of the key prosodic properties of language, influencing a whole range of means of expression, such as emphasis and emotion.

Test analysis of these phenomena was performed on the English word "processing". The recording of this expression, spoken by a native speaker (*lecturer*), is in .mp3 format and has duration of 985 ms. The *student* record is in the same format and has length of 944 ms. For this reason, both signals are fitted to the percentage x-axis. The first parameter analyzed is **duration**. The following graph in Fig. 2 shows the waveforms of the input signals and the division into voiced / unvoiced parts by a thick black line, where a high value indicates voiced parts and low value means unvoiced parts. It provides the *student* with feedback on the duration of individual sections in comparison with the *lecturer*.

Graphic information about the duration of individual parts is supplemented by verbal feedback to the student in the following way: *"The first section (unvoiced) is 4% shorter; the second section (voiced) is 4% shorter, the third section (unvoiced) is 2% longer, the fourth section (voiced) is 4% longer, the fifth section (unvoiced) is 2% shorter, the sixth section (voiced) is 3% longer."*

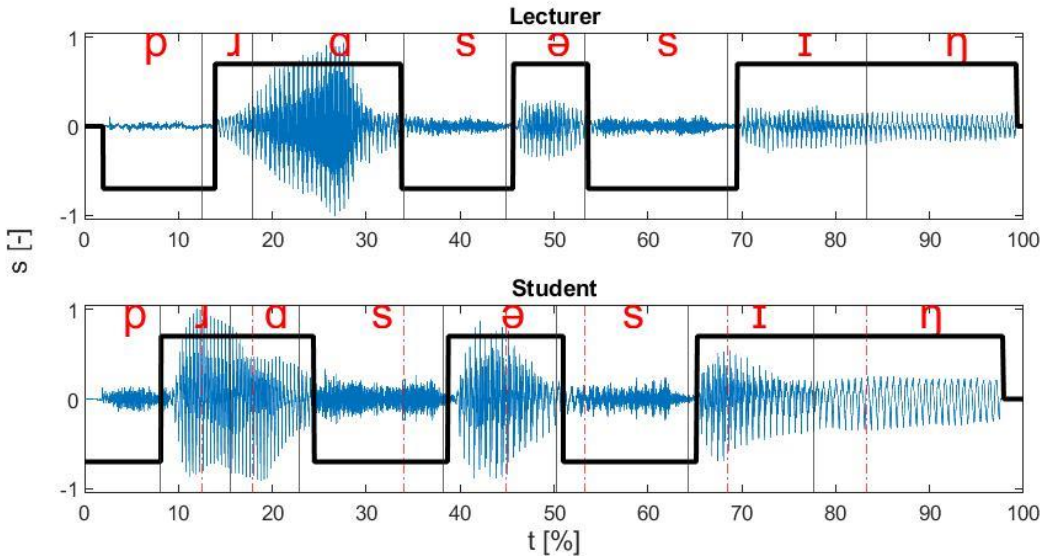

Figure 2: Duration of voiced and unvoiced segments

Energy the accent is another analyzed parameter. The graphs in Fig. 3 show the course of the normalized signal energy, centered around the mean value. This achieves a state where the volume of speech is irrelevant and only the emphasis on individual sounds or syllables is taken into account. The chart also serves as a simple feedback about the accent. Verbal feedback provides the student with information in the following way: *"Phoneme /ɹ/ should be unstressed; phoneme /ɪ/ should be unstressed."*

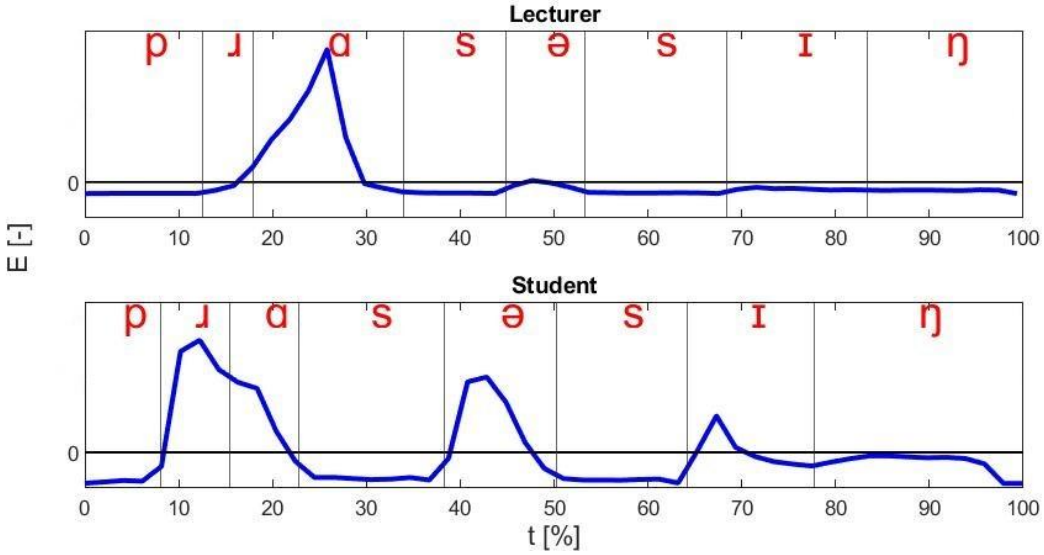

Figure 3: Energy (accent)

The third pronunciation parameter analyzed is pitch. The graphs in Fig. 4 show course of normalized tone and it gives key information about the locations of the raises or drops of the voice. It can be seen that the *lecturer's* signal shows a voice raise around 14% and a voice drop near 78%, yet the *student's* record shows only one voice drop around 44%. The verbal feedback here is: *"Voice rise missing at the /ɹ/ phoneme; voice fall missing at the /ɪ/ phoneme; voice fall at the /ə/ phoneme is incorrect."*

CONCLUSION

English places great emphasis on correct pronunciation and intelligibility of speech. However, pronunciation training is often neglected in language learning. The presented program therefore has great potential for increasing the ability of correct pronunciation, both in the form of an integrated module for dictionaries and in the form of a separate teaching tool. The goal of further development will be the extension of other methods of real-time pronunciation control.

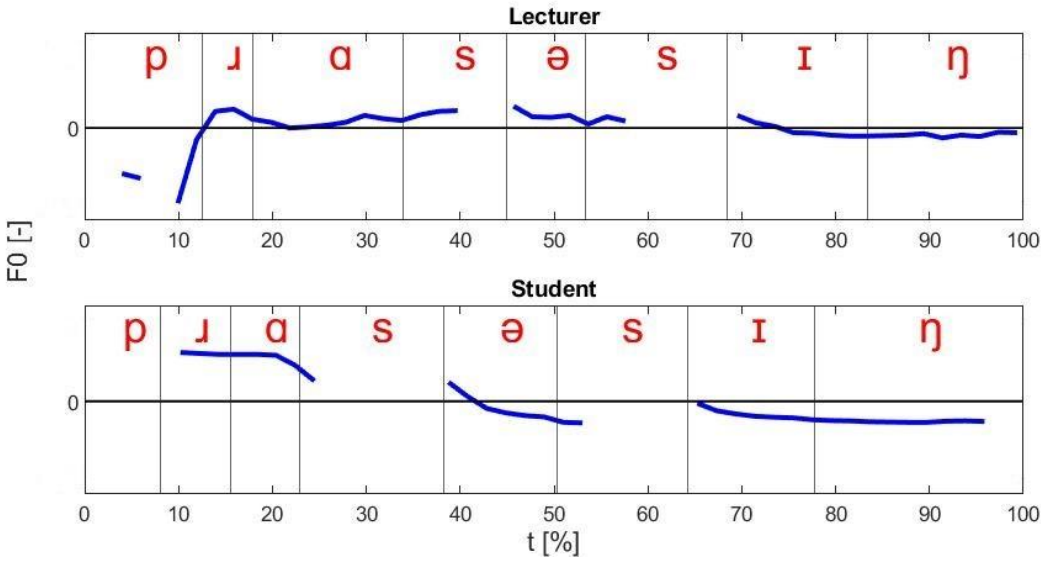

Figure 4: Pitch (melody)

ACKNOWLEDGMENT

Research described in this paper was supported by the Brno University of Technology Internal Grant Agency under project no. FEKT-S-20-6361.

REFERENCES

[1] Statistika & My | Magazín Českého statistického úřadu [online] from https://www.statistikaamy.cz/2017/10/17/ctyri-z-peti-cechuse-domluvi-cizi-reci/

[2] A. R. James, "The teaching of pronunciation," *TESOL Quarterly*, vol. 14, no. 2, pp. 246–250, 1980.

[3] C. Agarwal and P. Chakraborty, "A review of tools and techniques for computer aided pronunciation training (CAPT) in English," *Education and Information Technologies*, vol. 24, no. 6, pp. 3731–3743, 2019.

[4] M. F. Ürün, "Integration of technology into language teaching: A comparative review study," *Journal of Language Teaching and Research*, vol. 7, no. 1, pp. 76–87, 2016.

[5] PROZODIE V POPISU ZVUKOVÉ STAVBY JAZYKA | Nový encyklopedický slovník češtiny [online] from https://www.czechency.org

[6] R. Delmonte, "Prosodic tools for language learning," *International Journal of Speech Technology*, vol. 12, no. 4, pp. 161–184, 2009.

[7] J. Malucha, "Computer based evaluation of speech voicing for training English pronunciation," 29th Telecommunications Forum (TELFOR), Belgrade, 2021, pp. 1–4.

[8] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*. London: Prentice Hall, 2011.

[9] M. Sigmund, "Statistical analysis of fundamental frequency based features in speech under stress," *Information Technology and Control*, vol. 42, no. 3, pp. 286–291, 2013.

[10] M. Sigmund, "Spectral analysis of speech under stress," *International Journal of Computer Science and Network Security*, vol. 7, no. 4, pp. 170–172, 2007.

[11] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 5, pp. 353–362, 1974.