Journal of ELECTRICAL ENGINEERING, VOL 73(2022), NO5, 359–362

sciendo

COMMUNICATIONS

Comparison of methods for determining speech voicing based on tests performed on paired consonants and continuous speech

Jan Malucha, Milan Sigmund[1]

Voicing is an important phonetic characteristic of speech. Each phoneme belongs to a group of either voiced or unvoiced sounds. We investigated and compared the performance of five algorithms widely used to estimate speech voicing. All algorithms were implemented in Matlab and tested on both short consonants and continuous speech. Phonetically paired consonants (voiced *vs* unvoiced) and parts of read speech from audio books were used in the experiments. The tuned harmonics-to-noise ratio method gave the best results in both situations, *ie* for consonants and continuous speech. Using this method, the overall voicing of Czech, Polish, Hungarian and English was investigated. Hungarian speech showed the highest proportion of voiced parts, approx. 75 %. In other languages, the proportion of voiced parts was around 70 %.

K e y w o r d s: voiced and unvoiced speech, phonetics, evaluation of methods, statistical voicing characteristics of languages

1 Introduction

The standard transcription system widely used in linguistics is the International Phonetic Alphabet (IPA) [1]. An updated IPA including acoustic pronunciation samples can be found on the interactive website [2]. In addition, there are specially designed alphabets. For Czech and other slavic languages, the special slavic phonetic alphabet (SPA) is most suitable [3]. Linguists studying native american languages created the american phonetic alphabet (APA) with its own alternative symbols [4]. Interestingly, as an equivalent to the IPA symbols [ʧ, ɲ, ʃ, ʒ] the APA contains the symbols [č, ň, š, ž], which are also typical for the Czech language. To handle some special sounds not covered by the original IPA, a symbol set known as Extensions to the IPA (ExtIPA) was assembled. The latest revision of the ExtIPA is described in [5].

In European languages, all vowels (with pronunciation short and long) are typical representatives of voiced sounds. Also, vowel-like sounds, called semivowels, are voiced, *eg* /l/, /r/. On the other hand, there are paired phonemes in the consonant group (so called minimal pairs), for example paired stops such as voiced /g/, /d/ *eg* unvoiced /k/, /t/ or paired fricatives such as voiced /v/, /z/ *eg* . unvoiced /f/, /s/. Paired voiced phonemes are not always strictly pronounced as voiced. Due to coarticulation, they can change into their unvoiced counterparts. These phonetic variations are a common phenomenon in the Czech language.

There are several approaches to automatically distinguish between voiced and unvoiced speech segments. From the signal processing point of view, voicing can be defined as the presence of a fundamental tone. The physical energy source of this tone is the airflow from the respiratory system, which is further modulated by the vocal cords – the speed of their movement with the change in the width of the gap between them (lat. glottis) determines the voice's varying frequency, *ie* pitch. The resulting signal is then referred to as the glottal stream, which is quasi-periodic and contains distinctive pulses [6]. It is further filtered by the speech organ cavities and converted into individual voiced phonemes in the final speech signal. The presence of glottal pulses can be measured non-invasively using a special medical device – the electroglottograph, and it can be understood as an objective and highly accurate information about the sonority of a given segment of speech. Computational detection of glottal pulses in speech recording is dealt with by [7] or [8]. However, since these are in general computationally demanding methods, alternative methods suitable also for very fast, low-power or real-time applications are being developed.

The simplest methods work in the time domain. The short-time energy (STE) method is presented, for example, in [9]. The zero-crossing rate (ZCR) method based on the relative number of signal zero crossings is described in [10]. Several methods are based on the autocorrelation function and its modifications. The study [11] deals with the harmonics-to-noise ratio (HNR) method, which obtains information about voicing from autocorrelation peaks. In the method described in [12], voicing is determined by the ratio of autocorrelation function peaks of speech signal after center clipping. Several methods deal with determining the voicing from the spectral characteristics of a signal. The method in [13] converts the short-time spectrum of the signal into a normalized probability distribution, where each subcarrier is understood as a random variable, and voicing is evaluated according to

[1] Brno University of Technology, FEEC, Department of Radio Electronics, Technicka 12, 616 00 Brno, Czechia, 203286@vut.cz, sigmund@vut.cz

https://doi.org/10.2478/jee-2022-0049, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X
© This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (http: //creativecommons.org/licenses/by-nc-nd/3.0/).

probability concentrations. Other method based on cepstrum can be found in 14]. However, no method is considered by experts in the field to be clearly the best in terms of reliability and accuracy.

We briefly describe the individual algorithms used to estimate the voiced parts in speech signal and present the experimental results to compare the success rates of the tested method in selected phonetic situations. Further, the same methods are compared on the continuous speech signal of four different languages.

2 Used methods

1) Short-time energy (STE)

$$E_{\mathrm{ST}} = \sum_{n=0}^{N-1} [x(n)]^2 , \tag{1}$$

where N is the number of signal samples $x(n)$ in one segment. Thanks to the noisy and low-energy nature of unvoiced phonemes, the voicing of a segment can be determined easily using a variable threshold. However, this method is not sufficiently robust to signal dynamics and a varying SNR level.

2) Zero crossing rate (ZCR)

$$R_{\mathrm{ZC}} = \sum_{n=0}^{N-1} \big|\mathrm{sgn}\,[x(n)] - \mathrm{sgn}\,[x(n-1)]\big|, \tag{2}$$

where N is the number of signal samples $x(n)$ in one segment and sgn stands for the signum function. This method roughly estimates the frequency distribution of the segment. In general, voiceless phonemes have a character of high-frequency noise. Therefore, a variable threshold can be used to determine the segment's voicing. Unfortunately, this method is also sensitive to SNR and is not very suitable for determining silent regions.

3) Harmonic-to-noise ratio (HNR)

$$R_{\mathrm{HN}} = 10 \log_{10} \frac{r_2}{1 - r_2}, \tag{3}$$

where r_2 is normalized value of the second peak in the autocorrelation function. This method is based on the rate of signal harmonicity can be expressed as a ratio of normalized autocorrelation peak versus its complement. The ratio logarithm can be compared with a variable threshold to determine voicing of the segment. This method requires very accurate peak picking.

4) Spectral entropy (ENT)

$$H = -\sum_{x \in X} x_i \log_2 x_i, \quad x_i = \frac{X_i}{\sum_{i=1}^{N} X_i}. \tag{4,5}$$

Here, x_i is the segment's spectral probability function, X_i represents the energy of i-th frequency component and N is the number of frequency components.

The segment spectrum is first computed using FFT. Then the spectrum is normalized and converted into a function, which can be treated as a classic spectral probability distribution, in other words, we obtain a function, where each subcarrier represents a random process.

Entropy grows with increasing spectral noise. A high concentration of energy around one subcarrier indicates greater orderliness of the signal and thus presence of voicing. We can again choose a threshold to compare the value of entropy with. To increase accuracy, it is possible to modify the algorithm by dividing the full-band spectrum into sub-bands for separate entropy calculations [15].

5) Center clipping (CC)

At first, a clipping level for each i-th segment is calculated

$$C_{\mathrm{L}i} = k \min\{|MAX_{i-1}||MAX_{i+1}|\}, \tag{6}$$

where k is the so-called reduction factor (typical between 0.6 to 0.8) and MAX_{i-1} and MAX_{i+1} are the maximum values in adjacent segments $i-1$ and $i+1$. The speech signal is then clipped and normalized in each segment to 3 levels as follows

$$\hat{x}_i(n) = \begin{cases} +1 & \text{for } x_i(n) > C_{\mathrm{L}}\,, \\ 0 & \text{for } -C_{\mathrm{L}} \le x_i(n) \le C_{\mathrm{L}}\,, \\ -1 & \text{for } x_i(n) < -C_{\mathrm{L}}\,. \end{cases} \tag{7}$$

This means that all signal samples with an amplitude between $\pm C_{\mathrm{L}}$ are set to zero and remaining amplitudes are set to either $+1$ or -1. A standard autocorrelation function is obtained from the normalized signal. This method evaluates voicing based on the ratio of its first and following significant peak. The entire algorithm, including numerical parameters, is presented, for example, in [16].

3 Experimental testing of phonetic situations

The main goal of the experimental testing was to briefly compare the ability of the selected methods to identify voiced and unvoiced segments of the speech signal. Glottal pulses detected by the well-known program Praat [17] were considered as a reference for the correct detection of voiced speech. Detail shape of real glottal pulses can be seen, for example, in [18].

In the first series of experiments, the methods were compared on some special phonetic situations that cannot be easily verified by ear. Special situations here mean, on the one hand, the phenomenon of fading long vowels (usually at the ends of Czech words) and, on the other hand, the distinction of the so-called minimal pairs of voiced and unvoiced consonants. Experiments were carried out on real recordings of male voice in following speech signals: fading of the phoneme /a/ twice at the end of the words of the recording of the Czech phrase "škola volá", and distinguishing the voicing of phonemes in recordings of the minimal pairs /p/ - /b/ (occlusive), /dz/ - /dž/ (semiocclusive) and /f/ - /v/ (constrictive). All recordings in these experiments were made using the laptop's built-in microphone, which appears to be one of the most

Table 1. Comparison of investigated methods on phonemes

	Correct identification (%)				
	STE	ZCR	HNR	ENT	CC
Fading	79.49	89.74	84.62	87.18	87.18
Occlusive	92.68	85.37	92.68	87.81	87.81
Semiocclusive	67.65	85.29	97.06	55.88	82.35
Constrictive	95.83	95.83	91.67	81.25	91.67
Mean	83.13	88.95	91.39	76.78	87.18

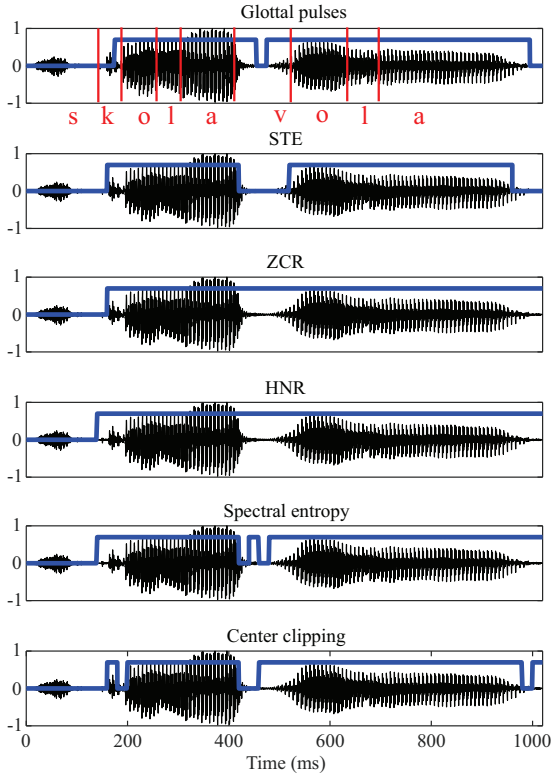

Fig. 1. Comparison of methods on the czech phrase "škola volá", where positive value of the blue lines indicates a voiced speech and zero value shows unvoiced or(almost)silent parts

common interfaces in the average user's environment. The basis for statistical comparison of all evaluated methods are the binary voicing curves along the time axis of the recordings for each method, as illustrated in Fig. 1 (blue lines).

The voicing curves obtained by the programmed algorithms were then compared to the corresponding reference curve from [17] to determine the percentage accuracy of each tested method in the given phonetic situations. It is necessary to point out that the variable thresholds in the individual methods were tuned up automatically to achieve the highest possible percentage similarity with the reference method of glottal pulses. The results achieved are summarized in Tab. 1.

Subsequently, the ability to identify voicing was tested on recordings of continuous speech on a group of professional recordings of Czech, Polish, Hungarian and English read speech with a length of 10 minutes, again spoken by male voices (audio books). All the texts are extracts from

Table 2. Comparison of investigated methods on continuous speech in different languages

	Correct identification (%)				
	STE	ZCR	HNR	ENT	CC
Czech	78.90	75.75	77.63	77.24	50.30
Polish	76.86	73.76	78.29	78.36	56.39
English	72.10	77.80	83.60	82.64	66.71
Hungarian	80.64	77.19	82.00	82.35	65.37
Mean	77.06	76.11	80.34	80.11	59.30

classical literature. Long stretches of silence and breathing were manually cut out from all recordings, so only parts of voiced and unvoiced speech were left for further processing. The results are presented in Tab. 2.

Obviously, the harmonics-to-noise ratio (HNR) method can be tuned to give the best overall results, which is shown in both tables. At the same time, it can be stated that center clipping is the weakest method when analyzing continuous speech; this may be due to the low robustness of the peak-picking algorithm. Some of the methods themselves are susceptible to various acoustic phenomena, such as noise or non-zero energy after phoneme fading, which can further deteriorate their performance to varying degree.

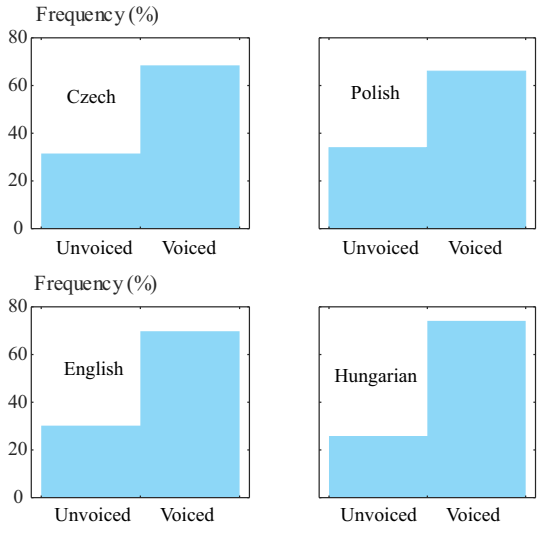

Fig. 2. Distribution of voicing in various languages derived from the presence of glottal pulses in speech signal

Interestingly, the overall ratio of voiced and unvoiced segments is very similar in all four languages, as is shown in Fig. 2. In addition, this has been also proven for the Swedish and German languages in further testing. This fact alone may provide a basic indicative measure of voicing detection accuracy for various methods and algorithms while analyzing longer continuous streams of speech in European languages. The binary histograms in Fig. 3 graphically compare the percentage of unvoiced/voiced relationships estimated by two selected methods, *ie* HNR and ENT, against the glottal pulse method. Both methods HNR and ENT return the best average results for continuous speech and their results differ

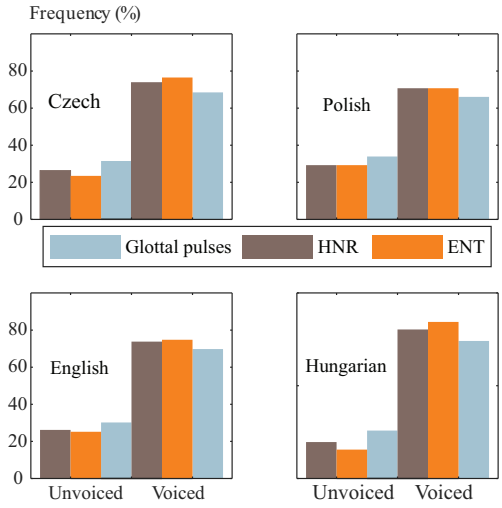

Fig. 3. Comparison of voicing distributions in various languages estimated by glottal pulses, HNR, and ENT

very little (see Tab. 2). It is evident, that both methods have a slight tendency more often to classify a not sufficiently clear speech segments as voiced, even if they are voiced. This may be since both methods sometimes fail to respond in time to the phoneme fading, where the reference algorithm no longer detects exciting glottal pulses, which can be clearly seen in Fig. 1.

4 Conclusion

Voicing is one of the basic features of speech signals and it is of great importance for several speech processing applications, primarily in language learning including synthetic speech [19]. Some special approaches in speech analysis are based only on voiced segments, for example stress monitoring [20], detection of deception [21], signaling of some neurodegenerative diseases such as Parkinson's disease [22] or Alzheimer's dementia [23].

In future work, we will test the algorithms under acoustic conditions that may affect their reliability, such as presence of non-stationary noise [24] and vocal effort fluctuations [25]. Our goal is to find an optimal fusion of effective algorithms for the correct classification of voiced and unvoiced sounds in continuous Czech speech.

Acknowledgements

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

References

[1] *International Phonetic Association, Handbook of the International Phonetic Association* A Guide to the Use of the International Phonetic Alphabet, Cambridge University Press, 1999.

[2] https://www.internationalphoneticassociation.org/IPAcharts/inter_chart_2018/IPA_2018.html.

[3] R. Sussex and P. Cubberley, *The Slavic Languages*, Cambridge University Press, 2006.

[4] D. Odden, *Introducing Phonology*, New York, Cambridge University Press, 2005.

[5] M. J. Ball, S. J. Howard, and K. Miller, "Revisions to the ExtIPA,", *Journal of the International Phonetic Association*, pp. 156–164.

[6] M. Sigmund, A. Prokes, and Z. Brabec, "Statistical Analysis of Glottal Pulses in Speech under Psychological Stress,", *16th European Signal Processing Conference (EUSIPCO)*, pp. 1–5, 2008.

[7] M. M. Sondhi, "Measurement of the Glottal Waveform,", *The Journal of the Acoustical Society of America*, no. 1, pp. 228–232, 1975.

[8] R. L. Miller, "Nature of the Vocal Cord Wave,", *The Journal of the Acoustical Society of America*, no. 6, pp. 667–677, 1959.

[9] J. Psutka, L. Müller, J. Matoušek, and V. Radová, *Mluvíme s počítačem česky*, Prague: Academia, 2006, (in Czech).

[10] R. G. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, "Voiced/unvoiced Decision for Speech Signals Based on Zero Crossing Rate and Energy,", *Advanced Techniques in Computing Sciences and Software Engineering*, Dordrecht: Springer, pp. 279–282, 2010.

[11] J. Heranová, *Harmonicita jako možný indiktor hranic mezi segmenty v češtině*, Praha, Univerzita Karlova, Filozofická fakulta, Fonetický ústav, 2010, (in Czech).

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

[13] H. Misra, S. Ikbal, H. Bourlard, and H. Hermansky, "Spectral Entropy Based Feature for Robust ASR,", *Int. Conference on Acoustics, Speech, and Signal Processing*, pp. I193–I196, 2004.

[14] P. J. Murphy and O. O. Akande, "Noise Estimation in Voice Signals Using Short-Term Cepstral Analysis,", *The Journal of the Acoustical Society of America*, no. 3, pp. 1679–1690, 2007.

[15] Y. Wei, Y. Zeng, and C. Li, "Single-Channel Speech Enhancement Based on Subband Spectral Entropy,", *Journal of the Audio Engineering Society*, no. 3, pp. 100–113, 2018.

[16] M. Sigmund, "Statistical Analysis of Fundamental Frequency Based Features in Speech Under Stress", *Information Technology and Control*, no. 3, pp. 286–291, 2013.

[17] P. Boersma and D. Weenink, "Praat: doing phonetics by computer", http://www.praat.org, 2022.

[18] M. Stanek and M. Sigmund, "Psychological Stress Detection in Speech Using Return-to-Opening Phase Ratios in Glottis,", *Elektronika ir Elektrotechnika*, no. 5, pp. 59–63, 2015.

[19] J. Pribil, A. Pribilova, and J. Matousek, "Evaluation of Synthetic Speech Quality by Statistical Analysis of Voiced and Unvoiced Part Durations,", *Int. Conference on Telecommunications and Signal Processing (TSP)*, pp. 1–4, 2018.

[20] M. Sigmund and T. Dostal, "Analysis of Emotional Stress in Speech,", *International Conference on Artificial Intelligence and Applications, Innsbruck*, pp. 317–322, 2004.

[21] S. Sondhi, R. Vijay, M. Khan, and A. K. Salhan, "Voice Analysis for Detection of Deception,", *Int. Conference on Knowledge, Information and Creativity Support Systems*, pp. 1–6, 2016.

[22] A. Bayestehtashk, M. Asgari, I. Shafran, and J. McNames, "Fully Automated Assessment of the Severity of Parkinson's Disease from Speech,", *Computer Speech and Language*, no. 1, pp. 172–185, 2015.

[23] M. L. B. Pulido, J. B. A. Hernández, M. A. F. Ballester, C. M. T. González, J. Mekyska, and Z. Smékal, "Alzheimer's Disease and Automatic Speech Analysis: A Review,", *Expert Systems with Applications*, pp. 1–19, 2020.

[24] P. Zelinka and M. Sigmund, "Hierarchical Classification Tree Modeling of Nonstationary Noise for Robust Speech Recognition", *Information Technology and Control*, no. 3, pp. 202–210, 2010.

[25] P. Zelinka and M. Sigmund, "Automatic Vocal Effort Detection for Reliable Speech Recognition", *Int. Workshop on Machine Learning for Signal Processing*, Kittila, pp. 349–354, 2010.

Received 8 September 2022