Spectral approximations for sibilant classification

Adam Stráník, Roman Čmejla

Faculty of Electrical Engineering, Czech Technical University in Prague
Prague 6, Technická 2, 166 27
stranada@fel.cvut.cz; cmejla@fel.cvut.cz

Abstract — **This paper deals with an analysis of power spectra of prolonged Czech sibilants /s/, /sh/, /z/, /zh/. Several approaches based on an approximation of the power spectral density (PSD) in several frequency bands were used. The experiments performed used a linear and a quadratic approximation in the frequency bands 1-3 kHz and 3-6 kHz, a cubic approximation in the frequency band 1-6 kHz and the Bézier approximation of a PSD which uses frequencies 0.3, 2, 6, 10 kHz as sampling points. The resulting parameters were used for classification of sibilants with a precision of 97.4% and with a precision of 99.4% for classification of the place of articulation.**

I. INTRODUCTION

Sibilants are a subset of fricative consonants that are produced by the narrowing a supraglottal part of the vocal tract and such constriction generates a typical turbulence noise. In the case of the consonant /s/ and /z/ the narrowed part occurs in the mouth and is created by an elevation of the tongue upwards while the axial sides of the tongue are clamped to the upper teeth ridge. The slot remains between the dorsum tip of the tongue and the front part of the alveola. The lip slot is often quite narrow. These two sibilants are called alveolar sibilants. In the case of /sh/ and /zh/ the narrowed part occurs in the mouth as well, but the main constriction is situated deeper in the mouth cavity between the dorsum of the tongue and the palate. According to the place of articulation these two sibilants are called palato-alveolar. The only difference in pronunciation between /s/, /z/ and /sh/, /zh/ is the following: /s/ and /sh/ are unvoiced whereas /z/ and /zh/ are their voiced opposites. The spectrum of the second pair contains energy in frequencies corresponding to the fundamental frequency [1], [2], [3].

It is proven that moving the place of articulation deeper into the oral cavity (i.e. for palato-alveolar sibilants) results in lowering the spectral energy in the signal, for example [2], [3], [4], [5].

It has been shown that locus equations, spectral slopes or another approximation of power spectrum can be used to determine the place of articulation of sibilants (alveolar vs. palato-alveolar); however, there is no unified opinion about the most suitable approach. [6] used the locus equation, especially the slope and y-intercept of F2 (second formant of preceding vowel) onset, to determine the place of articulation with great results – the accuracy of this approach is 87,1%. [7] computes spectral slope for all sibilants in the frequency range between 11-16.95 kHz, in addition to the unvoiced sibilants from the frequency corresponding to the maximum amplitude in the power spectrum to the 16.95 kHz, for voiced sibilants from each spectral maxima up to 16.95 kHz. In [7] there is also mention of several frequency bands used by other researchers: 0.05-16.95 kHz, 0.05-10 kHz, 0.05-5 kHz, 0.2-16.95 kHz, 0.2-10 kHz and 0.1-6 kHz. The lower bound 50 Hz, 100 Hz and 200 Hz should suppress the effect of voicing, the upper bounds are set mainly due to aliasing filters used for experiments. [4] computes two spectral slopes for every sibilant: low-frequency band slope in 0-2.5 kHz and high-frequency band slope in 2.5-8 kHz. [4] also describes another method suitable to determine frequency bands in which the spectral slopes will be computed – the method is based on the main spectral peak across talkers and sibilants; the spectral peaks differ for alveolars (8 kHz) and palato-alveolar (3.3 kHz).

II. DATABASE

The analyzed records were recorded in a soundproof booth with an Edirol R-09HR recorder with an Opus 55.09 MK II SC microphone. The following parameters were used: sample frequency fs=44.1 kHz, signal resolution 16 bits per sample, microphone plugged into Mic input, microphone gain set to L, input level 60 and phantom power turned on. All other recorder features were turned off.

All recorded subjects had the microphone positioned on the left side of their face at a distance of approx. 2 or 3 cm from the corner of their mouth. To avoid records with a direct impact of the air stream onto the microphone membrane, a hearing control was performed. If a direct impact of the air stream was detected, the recording was interrupted, the position of the microphone adjusted and the whole recording of the subject

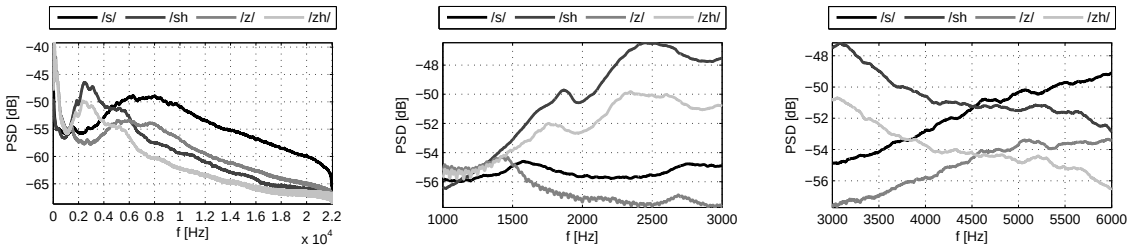

Figure 1: Power spectral density of all microsegments for sibilants /s/, /sh/, /z/, /zh/ in database in frequency range (a) 0-22050 Hz (22050 Hz is Nyquist frequency), (b) 1-3 kHz, (c) 3-6 kHz.

was repeated.

By now 39 subjects have been recorded — 36 males and 3 females. The mean age of recorded subjects is 30±8.7 years. During the recording every subject had to pronounce sustained sibilants /s/, /sh/, /z/, /zh/ for approx. 2 seconds each with a pause of approx. 2 seconds between sibilants and then had to make a fluent articulation changes in the following series: /s/→/sh/→/s/, /z/→/zh/→/z/, /s/→/z/→/s/, /sh/→/zh/→/sh/. Every sibilant in the series had to be sustained for approx. 2 seconds; the pause between series had to be approx. 2 seconds. The total number of analysed phonemes is 156 (39 subjects 4 sibilants each). Every record was segmented into microsegments the length of 20 ms with a 10 ms overlap. A Hamming window the length of 20 ms was applied to every microsegment. Microsegments corresponding to sustained sibilants /s/, /sh/, /z/, /zh/ were manually extracted from every record. The total number of microsegments of every sibilant is shown in Tab. 1.

Table 1: Number of microsegments of each sibilant in database.

sibilant	/s/	/sh/	/z/	/zh/
microsegments	7514	8394	7584	7641

III. EXPERIMENTS

Fig. 1(a) shows the power spectral density (PSD) of sibilants /s/, /sh/, /z/, /zh/. These PSDs were computed from the entire database, for example the PSD for /s/: every microsegment of /s/was zero-padded to the length of 44100, the DFT was applied and the power spectrum computed. The corresponding frequency lines from every power spectrum from each microsegment were averaged. As shown in Fig. 1(a), sibilants /s/ and /z/ have energy at higher frequencies than /sh/ and /zh/. The fundamental frequency of sibilants /z/ and /zh/ can also be determined.

The aim of our experiments was to find a regression of the power spectrum of sibilants which could separate every sibilant or group of sibilants (alveolar vs. palato-alveolar) with the best accuracy. According to Fig. 1(a) two frequency bands were used:

- **1-3 kHz** – energy of palato-alveolar sibilants is increasing sharply in this band whereas energy of alveolar sibilants either remains almost constant or decreases, see Fig. 1(b);
- **3-6 kHz** – energy of palato-alveolars is decreasing whereas energy of alveolars is increasing, see Fig. 1(c).

The linear approximation is represented by two parameters: $\mathrm{la_{L\text{-}U}}$ and $\mathrm{lb_{L\text{-}U}}$ obtained from line equation which best fits the underlying power spectrum in the specified frequency band

$$\mathrm{PSD_l} = \mathrm{la_{L\text{-}U}} f + \mathrm{lb_{L\text{-}U}}, \tag{1}$$

where $\mathrm{PSD_l}$ is a linear approximation of PSD, $\mathrm{la_{L\text{-}U}}$ [dB/Hz] is a slope and $\mathrm{lb_{L\text{-}U}}$ [dB] is an y-intercept of the regression line computed in the frequency band from L kHz to U kHz.

The quadratic approximation is represented by three parameters $\mathrm{qa_{L\text{-}U}}$, $\mathrm{qb_{L\text{-}U}}$ and $\mathrm{qc_{L\text{-}U}}$ corresponding to the quadratic equation describing a curve which best fits the underlying power spectrum

$$\mathrm{PSD_q} = \mathrm{qa_{L\text{-}U}} f^2 + \mathrm{qb_{L\text{-}U}} f + \mathrm{qc_{L\text{-}U}}, \tag{2}$$

where $\mathrm{PSD_q}$ is a quadratic approximation of PSD, $\mathrm{qa_{L\text{-}U}}$ [dB/Hz^2] is a quadratic coefficient, $\mathrm{qb_{L\text{-}U}}$ [dB/Hz] is a linear coefficient and $\mathrm{qc_{L\text{-}U}}$ [dB] is a constant term of a regression parabola which best fits the underlying PSD in the frequency band from L kHz to U kHz. Similar to the quadratic approximation the following annotation of coefficients is used in the cubic approximation

$$\mathrm{PSD_c} = \mathrm{ca_{L\text{-}U}} f^3 + \mathrm{cb_{L\text{-}U}} f^2 + \mathrm{cc_{L\text{-}U}} f + \mathrm{cd_{L\text{-}U}}, \tag{3}$$

where $\mathrm{PSD_c}$ is a cubic approximation of PSD, $\mathrm{ca_{L\text{-}U}}$ [dB/Hz^3] is a cubic coefficient, $\mathrm{cb_{L\text{-}U}}$

Table 2: Mean values and standard deviations of observed parameters across sibilants. Annotation of parameters is described in the text.

parameter	/s/	/sh/	/z/	/zh/
$la_{1\text{-}3}$	2.6e-4 ± 8.0e-4	5.2e-3 ± 2.7e-3	-1.6e-3 ± 1.0e-3	2.9e-3 ± 2.5e-3
$la_{3\text{-}6}$	2.0e-3 ± 1.2e-3	-1.4e-3 ± 8.7e-4	1.6e-3 ± 1.3e-3	-1.5e-3 ± 9.0e-4
$qa_{1\text{-}3}$	7.2e-8 ± 1.3e-6	-3.0e-6 ± 2.7e-6	9.1e-7 ± 1.2e-6	-1.8e-6 ± 3.0e-6
$qb_{1\text{-}3}$	-3.3e-5 ± 5.0e-3	1.7e-2 ± 1.1e-2	-5.3e-3 ± 4.8e-3	1.0e-2 ± 1.2e-2
$qa_{3\text{-}6}$	-3.1e-7 ± 1.4e-6	6.9e-7 ± 1.1e-6	-4.6e-7 ± 1.2e-6	4.2e-7 ± 1.2e-6
$qb_{3\text{-}6}$	4.8e-3 ± 1.2e-2	-7.7e-3 ± 1.0e-2	5.7e-3 ± 1.1e-2	-5.3e-3 ± 1.1e-2
$dBez_2$	1.2e0 ± 1.7e0	2.3e0 ± 1.7e0	-7.8e0 ± 2.6e0	-6.3e0 ± 2.9e0
$dBez_3$	2.5e0 ± 1.4e0	-1.6e0 ± 1.3e0	-2.2e0 ± 1.9e0	-5.2e0 ± 1.4e0
$dBez_4$	3.7e-1 ± 2.3e0	-5.0e0 ± 2.0e0	-1.8e0 ± 2.4e0	-5.3e0 ± 2.1e0
$ca_{1\text{-}6}$	-1.2e-10 ± 2.9e-10	5.5e-10 ± 2.8e-10	-2.4e-10 ± 2.6e-10	3.6e-10 ± 2.9e-10
$cb_{1\text{-}6}$	1.5e-6 ± 2.9e-6	-6.7e-6 ± 3.1e-6	3.1e-6 ± 2.5e-6	-4.4e-6 ± 3.2e-6
$cc_{1\text{-}6}$	-4.4e-3 ± 8.2e-3	2.5e-2 ± 1.1e-2	-1.1e-2 ± 7.1e-3	1.6e-2 ± 1.1e-2

[dB/Hz^2] is a quadratic coefficient, $cc_{L\text{-}U}$ [dB/Hz] is a linear coefficient and $cd_{L\text{-}U}$ [dB] is a constant term.

For Bézier approximation a cubic Bézier curve was used. Control points $\mathbf{B}(t)$ of the Bézier curve are obtained from four sampling points $\mathbf{P}_0$, $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$ as follows [8]

$$\mathbf{B}(t) = (1-t)^3\mathbf{P}_0 + 3(1-t)^2 t\mathbf{P}_1 + 3(1-t)t^2\mathbf{P}_2 + t^3\mathbf{P}_3, \quad (4)$$

where $t \in [0, 1]$, sampling points $\mathbf{P}$ are the PSD function values at frequencies described in Tab. 3 and $\mathbf{B}(t)$ are control points describing the approximation curve.

Table 3: Frequencies corresponding to the sampling points $\mathbf{P}$ used for Bézier approximation.

point	$\mathbf{P}_0$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_3$
f [kHz]	0.3	2	6	10

Regressions were performed on PSD computed from five consequential microsegments in one record (i.e. there is no microsegment leakage across records) which corresponds to 60 ms of frication noise. According to [3] the listener needs at least 50 ms to recognize an /s/.

In our experiments we would like to avoid parameters describing absolute amplitude (energy) values stored in the spectrum. Therefore the following parameters were excluded: y-intercept in the linear regression and the constant terms in the quadratic and the cubic regression. For the same reason the control points of Bézier approximation were transformed as follows

$$dBez_t = B_y(t) - B_y(t-1), \quad (5)$$

where $B_y(t)$ is the y-coordinate (i.e. decibels) of t-th control point $\mathbf{B}(t)$. The reason for this is to suppress a different speech volumes across recorded subjects.

In Tab. 2 there are mean values and standard deviations of observed parameters across sibilants.

IV. RESULTS

Classification experiments based on the features shown in Tab. 2 were performed by the WEKA Data Mining Tool [9]. All experiments were performed using 10-fold cross-validation.

The best algorithm is K-NN (K Nearest Neighbours) with the full set of twelve-dimensional feature vector. Setting K=1, the overall precision is 99.6%. Increasing K to achieve a more general classifier we obtain 94.1% precision for K=1000. This is still a very good result, but generally, this approach needs all 12 parameters to be computed and then a search in 12 dimensional feature space needs to be performed – this approach was found to be very time-consuming in real-time applications.

The way to speed up the K-NN classifier in real-time applications is to reduce the feature space from the current twelve dimensions to lower dimensions, the lower the better. The reduction of feature space was performed based on the gain ratio measure ([10], [11]). Gain ratios of analysed parameters are given in Tab. 4. Fig. 2 shows the dependence of the precision on the number of features. The result of this experiment is the following: a three-dimensional feature vector consisting of [$dBez_2$; $dBez_3$; la_{1-3}] provides sufficient precision 97.4% while K=5. Reducing the feature vector to two-dimensions resulted in a decrease of precision to 92.1%. Another way to look at K-NN is the choice of K – the higher the K the more suppression of noise effect in the input data, but more time-consuming classification. The dependence of precision on K is shown in Fig. 3. The best result is achieved for K=10.

V. CONCLUSION

This study demonstrates that there are relevant and measurable changes in the power spectrum along sibilants /s/, /sh/, /z/, /zh/. By continuous averaging of the power spectrum from five

Table 4: Gain ratios for analysed parameters. Parameters not shown in the table have zero gain ratio.

parameter	$dBez_2$	$dBez_3$	$la_{1\text{-}3}$	$cc_{1\text{-}6}$	$la_{3\text{-}6}$	$qb_{1\text{-}3}$	$dBez_4$	$qa_{1\text{-}3}$	$qb_{3\text{-}6}$
gain ratio	0.2644	0.2527	0.2381	0.2300	0.2230	0.1597	0.1419	0.1131	0.0732

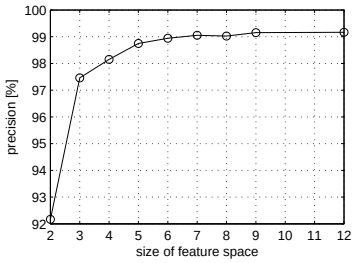

Figure 2: Dependence of precision on feature space for 5-NN algorithm. Features were removed according to their gain ratio shown in Tab. 4. In the first step all features with zero gain ratio were removed.

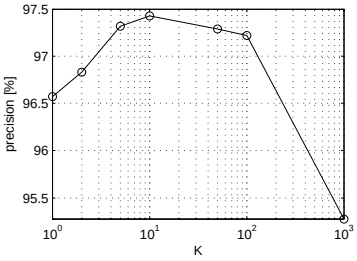

Figure 3: Dependence of precision on K (number of neighbours) in the feature space [$dBez_2$; $dBez_3$; la_{1-3}].

consequent microsegments and by using the parameters based on the linear approximation in frequency band 1-3 kHz and Bézier approximation with control points at frequencies 0.3, 2, 6, 10 kHz of this averaged spectrum we found that it is possible to classify this averaged spectrum into four classes corresponding to the sibilants /s/, /sh/, /z/ and /zh/. The presented classifier is based on the 10-NN (Nearest Neighbours) algorithm and is able to classify almost 97.4% of microsegments in the input database correctly. Reducing four classes corresponding to the sibilants to two classes corresponding to the place of articulation (alveolar vs. palato-alveolar) the precision of the 10-NN algorithm with three-dimensional feature space is 99.4%.

ACKNOWLEDGEMENT

This work has been supported by: **GACR102/08/H008** *Biological and Speech Signal Modelling*, **SGS10/180/OHK3/2T/13** *Assessment of voice and speech impairment*, **MSM6840770012** *Transdisciplinary Research in Biomedical Engineering II.*

REFERENCES

[1] PALKOVÁ, Zdena. (1994). *Fonetika a fonologie češtiny.* 367 pages.

[2] STEVENS, Kenneth N. (1998). *Acoustic Phonetics.* Cambridge, Massachusetts: The MIT Press, 607 pages.

[3] ABDELATTY ALI, Ahmed M; SPIEGEL, Jan Van der; MUELLER, Paul. (2001). *Acoustic-phonetic features for the automatic classification of fricatives.* In J. Acoust. Soc. Am. 109, 2217-2235.

[4] MANIWA, Kazumi; JONGMAN, Allard; WADE, Travis. (2009). *Acoustic characteristics of clearly spoken English fricatives.* In J. Acoust. Soc. Am. 125, 3962-3973.

[5] JONGMAN, Allard; WAYLAND, Ratree; WONG, Serena. (2000). *Acoustic characteristics of English fricatives.* In J. Acoust. Soc Am. 108, 1252-1263.

[6] SUSSMAN, Harvey M.; SHORE, Jadine (1996). *Locus equations as phonetic descriptors of consonantal place of articulation.* In Perception & Psychophysics, 58, 936-946.

[7] SHADLE, Christine H.; MAIR, Sheila J. *Quantifying Spectral Characteristics of Fricatives.* Available online on WWW: <http://www.asel.udel.edu/icslp/cdrom/vol3/951/a951.pdf>

[8] FARIN, Gerald. (1997). *Curves and surfaces for computer-aided geometric design (4 ed.).* Elsevier Science & Technology Books, ISBN 978-0-12249054-5

[9] HALL, M. et. al. (2009). *The WEKA Data Mining Software: An Update.* In SIGKDD Explorations. Volume 11, Issue 1.

[10] BURKE, J. P.; ZEIDLER, J. R.; RAO, B. D. (2005). *CINR Difference Analysis of Optimal Combining Versus Maximal Ratio Combining.* In IEEE Transactions on Wireless Communication. 4, p. 1-5.

[11] HARRIS Jr., Earl. (2001). *Information Gain Versus Gain Ratio: A Study of Split Method.* Available online on WWW: <http://www.mitre.org/work/tech_papers/tech_papers_01/harris_biases/harris_biases.pdf>

[12] HUGHES, George W; HALLE, Morris (1956). Spectral Properties of Fricative Consonants. In J. Acoust. Soc. Am. 28, 303-310.