Speech spectrum analyses for estimating operator functionality

O N Korsun1,2, V N Yurko1, S A Beyseev3, A S Naukenova3 and A K Tulekbaeva3

1 State Research Institute of Aviation Systems, 7 Viktorenko, Moscow 125319, Russia
2 Moscow Aviation Institute (National Research University), 4 Volokolamskoe Shosse, Moscow 125993, Russia
3 M. Auezov South Kazakhstan University, Shymkent 160012, Kazakhstan

marmotto@rambler.ru

Abstract. The paper deals with the problem of estimating human operator functional characteristics using frequency analysis of operator’s speech. The analysis is carried out using operator’s speech transfer function based on the classical definition of transfer function in the automatic control theory. The algorithms for calculating the operator’s speech transfer function are presented, the experimental results are discussed.

1. Introduction

In recent years, there has been an intensive development of man-machine interface tools in technology, such as modern audio technologies [1–3], neural networks [4] etc. In [5], the correlation between human operator physiological state and variations of his speech are investigated.

This work is devoted to the analysis of speech signal in the frequency domain for the purpose of estimating operator functionality. The analysis is based on the speech transfer function, which is introduced by analogy with the transfer function of the classical theory of automatic control [6].

2. Algorithm for calculating a frequency transfer function

Let us consider an audio record of a word and divide it into intervals of 20–40 ms. Every interval contains a sequence \( x(N) \), where \( N \) — the number of samples. Then the following operations are performed [3, 7]:

- signal pre-filtering
  \[ x(n) = x(n) - \alpha \cdot x(n-1), \alpha = 0.95; \]

- signal frame extraction
  \[ s_m(n) = x(m \cdot \Delta n + n), \quad 0 \leq n \leq N_{FFT} - 1, \]

where \( \Delta n = N_{FFT} - \varepsilon \cdot N_{FFT} \), \( N_{FFT} \) — length of fast Fourier transform in samples (frame length); \( \varepsilon \) — the ratio of the length of the overlap section to the length of the frame, \( 0 \leq \varepsilon \leq 0.5; \) \( m \) — frame number;

- frame weighting by Hanning window
\[ w_{H} = 0.5 \left( 1 - \cos \left( \frac{2\pi n}{N_{FFT} - 1} \right) \right) \] — Hanning window; 
\[ s_{m}(n) = w_{H}(n) \times s(n); \]

- calculating the fast Fourier transform and the square of its modulus, or estimating the spectral density

\[ X_{m}(k) = \text{FFT}\{s_{m}(n)\} = \sum_{n=0}^{N_{\text{FFT}}} s_{m}(n) \cdot e^{\frac{j \cdot 2\pi kn}{N_{\text{FFT}}}}, \]

\[ S_{m}(k) = |X_{m}(k)|^2 = \text{Re}(X_{m}(k))^2 + \text{Im}(X_{m}(k))^2, \]

- calculation of the average spectral density in the spectral band

\[ S(f_i) = \frac{1}{N_{\text{FFT}}} \sum_{k=k_{\text{min}}}^{k_{\text{max}}} S_{m}(k), i = 1, ..., N_{\text{frb}}, \]

where \( \Delta k = k_{\text{max}} - k_{\text{min}} + 1 \) — spectral band width; \( k_{\text{max}}, k_{\text{min}} \) — indices of the first and last frequency in the spectral band,

\[ k_{\text{min}} = \left\lfloor \frac{N_{\text{FFT}}}{f_s} \left( f_{\text{min}} + \frac{f_{\text{max}} - f_{\text{min}}}{N_{\text{frb}}} (i - 1) \right) \right\rfloor + 1, \quad i = 2 \ldots N_{\text{frb}}, \]

\[ k_{\text{max}} = \left\lfloor \frac{N_{\text{FFT}}}{f_s} \left( f_{\text{min}} + \frac{f_{\text{max}} - f_{\text{min}}}{N_{\text{frb}}} i \right) \right\rfloor, \quad i = 1 \ldots N_{\text{frb}}, \]

\[ f_i \] — frequency corresponding to the middle of the \( i \)-th band; \( f_s \) — sampling frequency;

Thus, on each time interval of 20 ... 40 ms duration, the speech signal is considered as a sequence of spectral density values averaged over the frequency band. This approach is based on the classical definition of the spectral density of the signal [6].

The algorithm described above is applied to each word realization, as a result, we obtain a matrix of dimension \( N_{\text{frb}} \times N_{\text{frb}} \), in which the number of rows \( N_{\text{frb}} \) is equal to the number of spectral bands, and the number of columns \( N_{\text{frb}} \) is the number of time quantization intervals. Normally the number of spectral bands \( N_{\text{frb}} = 35 \), the number of time intervals \( N_{\text{frb}} = 17 \).

Consider calculating the transfer function between two speakers to reveal the differences in frequency properties.

Let the first speaker say \( N \) realizations of a word.

The average matrix is calculated

\[ X_{D1} = \frac{1}{N} \sum_{i=1}^{N} \{x_{ij}\}_{i=1}^{N}; i = 1, ..., N_{\text{frb}}; j = 1, ..., N_{\text{frb}}, \]

where \( N_{\text{frb}} \) — is the number of time intervals into which the considered word record is divided; \( N_{\text{frb}} \) — the number of spectral bands.

The matrix for the second speaker \( X_{D2} \) is calculated similarly. If it is required to find changes in speech properties between different states of one speaker, then the \( X_{D2} \) matrix is calculated for the second state of this speaker.
When calculating the transfer function, we additionally carry out averaging for each frequency band over \( N \) time intervals. Finally, we get the formula for the modulus of the transfer function between the first and second speakers in decibels (dB):

\[
W_{12}(f_j) = 10 \log_{10} \frac{\frac{1}{N} \sum_{i=1}^{N} \{x_{ij}\}_{D1}}{\frac{1}{N} \sum_{i=1}^{N} \{x_{ij}\}_{D2}}.
\]

3. Dependence of the estimates of the speech transfer function on the structure and volume of speech material

To test the performance of the above algorithm, the following experiment was performed. Under normal conditions, the speaker pronounced the Russian words "пилотаж" (aerobatics), "масштаб" (scale), "навигация" (navigation) in 50 realizations of each. Note that the analysis of the above algorithm suggests that it does not depend on the language used. Indeed, in [7] for a similar algorithm using examples of Russian, Kazakh and English languages, it was shown that changing the language does not affect the accuracy of automatic word recognition.

The experiment was repeated after 4 hours, also under normal conditions. Based on the data obtained, the transfer functions were calculated separately for each word for all its implementations (Figure 1) As you can see, the values of the function basically correspond to a unity gain (level 0 dB), which indicates the constancy of the speaker properties. In this case, deviations of this do not exceed \( \pm 1.5 \) dB. Figure 1 also characterizes the degree of dependence of estimates on the structure of speech material: deviations of estimates calculated for individual words relative to the average do not exceed \( \pm 1 \) dB.

![Figure 1. Graph of the speech transfer function for the speaker D1 according to the speaker D1 in normal conditions with an interval of 4 hours between recordings.](image)

To establish the dependence of the estimates averaged over three words on the number of realizations, we compared the values of the function calculated over 50 realizations and taken as a standard with estimates for \( n = 3, 5, 10, 15, 20, 25 \) realizations. The standard deviations (s.d.) \( \sigma_i \) of each estimate from the standard were considered as a measure of the error (Table 1). The table shows that already with 3 realizations the s.d. of errors do not exceed 0.5 dB, which is sufficient to detect significant changes.

| \( N \) | 3   | 5   | 10  | 15  | 20  | 25  |
|--------|-----|-----|-----|-----|-----|-----|
| \( \sigma_i \), dB | 0.492 | 0.395 | 0.212 | 0.179 | 0.092 | 0.049 |
Let us further consider several examples characterizing the capabilities of the proposed speaker's transfer function for detecting changes in speech in various conditions. In these examples, to eliminate the influence of sample size, estimates are calculated for 50 realizations of each word.

4. Influence of noise-protective headphones on speech
This section deals with the transfer functions for speakers in normal conditions, and for the same speakers in noise protection headphones.

![Graph of the transfer function for the speaker D2 in noise-protective headphones.](image1)

**Figure 2.** Graph of the transfer function for the speaker D2 in noise-protective headphones.

![Graph of the transfer function for the D4 speaker in low frequency noise conditions, part 1.](image2)

**Figure 4.** Graph of the transfer function for the D4 speaker in low frequency noise conditions, part 1.

The resulting graphs show the effect of noise protection headphones on the speaker. It should be noted that the reactions of different speakers differ from each other, however, typical is the change in amplitude up to ±3 dB in the frequency range 6 ... 11 kHz.

5. Impact of noise on the state of a speaker over time
Consider the issue of changing the speaker's state during the experiment. To do this, a set of 150 words was divided into 4 parts, after which the transfer function for each part was calculated. The total time of the experiment was 2 minutes, the duration of one part was 0.5 minutes. During the experiment, the speaker was exposed to a low-frequency noise of 90 dB in the frequency band up to 2 kHz.
In general, the graphs for each part are similar, but the amplitude of the graphs of the 4th part is lower. This can be explained by the fact that over time the speaker gets used to the noise and the volume of his voice decreases, that is, the Lombard effect weakens (Lombard effect — the instinctive rise of the speaker voice volume in noisy environment [1]).

6. The influence of hearing defects on the frequency properties of speech
Let us consider the application of the proposed function to identify the speech features of a group of helicopter pilots who have an occupational hearing disorder — hearing loss. A speaker who had no diagnosed diseases of hearing and speech was taken as a reference speaker. The results for the three speakers in this group are shown in Figure 6-8.
Comparison with the transfer function of a speaker without hearing diseases (Figure 1, the spread is within $\pm 1.5 \, \text{dB}$), shows that the graphs of the transfer functions for this group of speakers differ in a greater variation of values: for speaker D5 from $-4$ to $3 \, \text{dB}$, for speaker D6 from $-5$ to $2 \, \text{dB}$, for speaker D7 from $-6$ to $4 \, \text{dB}$. Thus, the method considered in the article provides a good diagnostic feature for identifying speakers with hearing diseases based on their speech samples.

![Figure 8. Graph of the transfer function for the speaker D7](image)

7. Conclusion
The paper reports the following main results:
- an algorithm for calculating the speaker's speech transfer function, which enables to analyze changes in the speech frequency spectrum due to the external conditions or the speaker’s state;
- examples are presented that characterize the capabilities of the proposed speech transfer function to detect the state of the speaker external conditions or.

Acknowledgements
This work was supported by Russian Foundation for Basic Research (RFBR), project 18-08-01142-a.

References
[1] Rabiner L and Juang B H 1993 *Fundamentals of Speech Recognition* (Englewood Cliffs, New Jersey: Prentice Hall International, Inc.) p 507
[2] Bondaros Yu G, Ivanov A I and Tischenko A A 2010 Operator fatigue degree definition according his voice Lyapunov exponent *Herald of computer and information technologies* **6** 22–30
[3] Korsun O N and Poliev A V 2016 Automated definition of phonetically homogeneous sections of words in a natural language based on multiparameter optimization *Journal of Computer and Systems Sciences International* **55** (4) 609–18
[4] Evdokimenkov V N et al. 2015 The use of the neural network model of pilot control actions for their individually-adapted support *Journal of Computer and Systems Sciences International* **54** (4) 609–20
[5] Ivanov A I et al. 2014 Study of speech characteristics in hearing-impaired pilot speakers for creation of voice-activated system operating airborne equipment *Occupational Health and Industrial Ecology* **11** 40–5
[6] Methods of classical and modern theory of automatic control 2004 ed K A Pupkov and N D Egupov (Moscow: Izd. MGTU im. Bauman) p 656
[7] Korsun O N, Gabdrakhmanov A Sh, Mihajlov E I, Naheaev M Z and Tulekbaeva A K 2015 Algorithm for an automatic recognition of the speech commands, invariant to languages *Mekhatronika, Avtomatizatsiya, Upravlenie* **16** (9) 599–604