UDC 004.934

DOI: 10.18372/2310-5461.38.12858

O. Lavrynenko, Postgraduate
National Aviation University
E-mail: oleksandrlavrynenko@gmail.com

G. Konakhovych, D.Eng.Sc., prof.
National Aviation University

D. Bakhtiiarov, Postgraduate
National Aviation University

COMPARATIVE ANALYSIS OF SPEECH RECOGNITION ALGORITHMS IN UAV VOICE CONTROL SYSTEM

Introduction

Although automatic speech recognition systems have dramatically improved in recent decades, speech recognition accuracy still significantly degrades in noisy environments. While many algorithms have been developed to deal with this problem, they tend to be more efective in stationary noise such as white or pink noise than in the presence of more realistic degradations such as background music, background speech, and reverberation. At the same time, it is widely observed that the human auditory system retains relatively good performance in the same environments.

Problem statement

The goal of this thesis is to use mathematical representations that are motivated by human auditory processing to improve the accuracy of automatic speech recognition systems. Throughout this work we propose a number of signal processing algorithms that are motivated by these observations and can be realized in a computationally efficient fashion using real-time online processing. We demonstrate that these approaches are efficient in improving speech recognition accuracy in the presence of various types of noisy and reverberant environments.

Comparative Analysis of Speech Recognition Algorithms in UAV Voice Control System

The Frequency scales describe how the physical frequency of an incoming signal is related to the representation of that frequency by the human auditory system. In general, the peripheral auditory system can be modeled as a bank of bandpass filters, of approximately constant bandwidth at low frequencies and of a bandwidth that increases in rough proportion to frequency at higher frequencies. Because di□erent psychoacoustical techniques provide somewhat different estimates of the bandwidth of the auditory filters, several different frequency scales have been developed to fit the psychophysical data. Some of the widely used frequency scales include the MEL scale, the BARK scale, and the ERB (Equivalent rectangular bandwidth) scale. The popular Mel Frequency Cepstral Coefficients (MFCCs) incorporate the MEL scale, which is represented by the following equation:

$$Mel(f) = 2595 \log\left(1 + \frac{f}{700}\right).$$

The MEL scale that was proposed by Stevens describes how a listener judges the distance between pitches (Fig. 1).

The reference point is obtained by defining a 1000 Hz tone 40 dB above the listener's threshold to be 1000 mels.

Another frequency scale, called the Bark scale, was proposed by Zwicker:

$$Bark(f) = 13 \arctan(0{,}00076 f) + 3.5 \arctan\left(\frac{f}{7500}\right)^2.$$

Frequency relation is based on a similar transformation given by Schroeder:

$$\Omega(f) = 6 \ln\left(\frac{f}{600} + \left(\frac{f}{600}\right)^{0,5}\right).$$

More recently, Moore and Glasberg proposed the ERB (Equivalent Rectangular Bandwidth) scale modifying Zwicker's loudness model.

The ERB scale is a measure that gives an approximation to the bandwidth of filters in human hearing using rectangular bandpass filters; several different approximations of the ERB scale exist.

The following is one of such approximations relating the ERB and the frequency *f*:

$$ERB(f) = 11.17 \log\left(1 + \frac{46.065 f}{f + 14678.49}\right).$$

Fig. 1 compares the three different frequency scales in the range between 100 Hz and 8000 Hz. It can be seen that they describe very similar relationships between frequency and its representation by the auditory system [1].

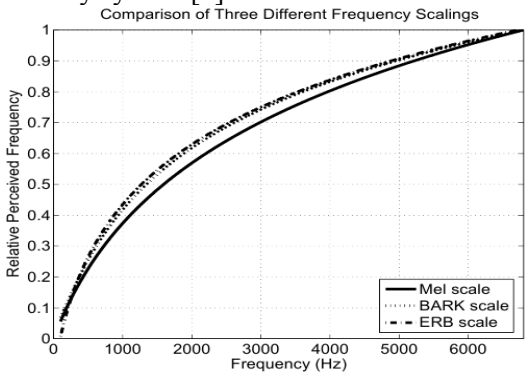

Fig. 1. Comparison of the MEL, Bark, and ERB frequency scales

Auditory nonlinearity is related to how humans process intensity and perceive loudness. The most direct characterization of the auditory nonlinearity is through the use of physiological measurements of the the average firing rates of fibers of the auditory nerve, measured as a function of the intensity of a pure-tone input signal at a specified frequency. As shown in Fig. 2, this relationship is characterized by an auditory threshold and a saturation point. The curves in Fig. 2 are obtained using the auditory model developed by Heinz [2].

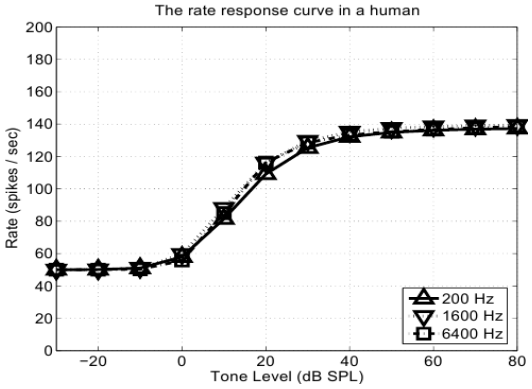

Fig. 2. The rate-intensity function of the human auditory system as predicted by the model of Heinz et al. for the auditory-nerve response to sound

Another way of representing auditory nonlinearity is based on psychophysics. One of the well-known psychophysical rules is Steven's power law, which relates intensity and perceived loudness in a hearing experiment by fitting data from multiple observers in a subjective magnitude estimation experiment using a power function:

$$L = \left(\frac{I}{I_0}\right)^3.$$

Another common relationship used to relate intensity to loudness in hearing is the logarithmic curve, which was originally proposed by Fechner to relate the intensity-discrimination results of Weber to a psychophysical transfer function. MFCC features, for example, use a logarithmic function to relate input intensity to putative loudness, and the definition of sound pressure level (SPL) is also based on the logarithmic transformation:

$$L_p = 20 \log_{10}\left(\frac{p_{rms}}{p_{ref}}\right).$$

The commonly-used value for the reference pressure p_{ref} is $20\mu Pa$, which was once considered to be the threshold of human hearing, when the definition was first established [3].

In Fig. 3, we compare these nonlinearities. In addition to the nonlinearities we included another power-law nonlinearity which is an approximation to the physiological model of Heinz et al. between 0 and 50 dB SPL in the Minimum Mean Square Error (MMSE) sense. In this approximation, the estimated power coefficient is around 1/10.

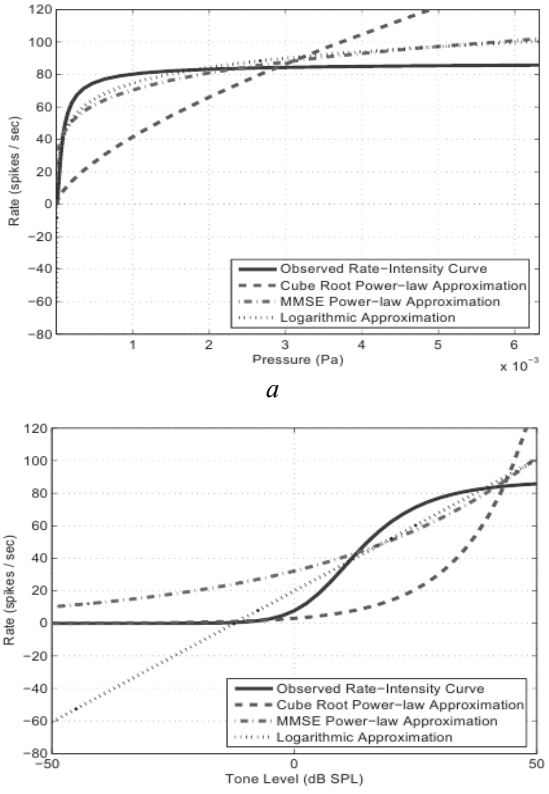

Fig. 3. Comparison of the cube-root power law nonlinearity, the MMSE power-law nonlinearity, and logarithmic nonlinearity. Plots are shown using two different intensity scales: pressure expressed directly in P_a (upper panel) and pressure after the log transformation in dB SPL (lower panel)

In Fig. 3, *a* we compare these curves as a function of sound pressure directly as measured in P_a. In

this figure, with the exception of the cube power root, all three curves are very similar.

Nevertheless, if we plot the curves using the logarithmic scale (dB SPL) to represent sound pressure level, we can observe a significant difference between the power-law nonlinearity and the logarithmic nonlinearity in the region below the auditory threshold. This difference plays an important role for robust speech recognition [4].

The most widely used forms of feature extraction are Mel Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP). MFCC processing begins with pre-emphasis, typically using a first-order high-pass filter (Fig. 4).

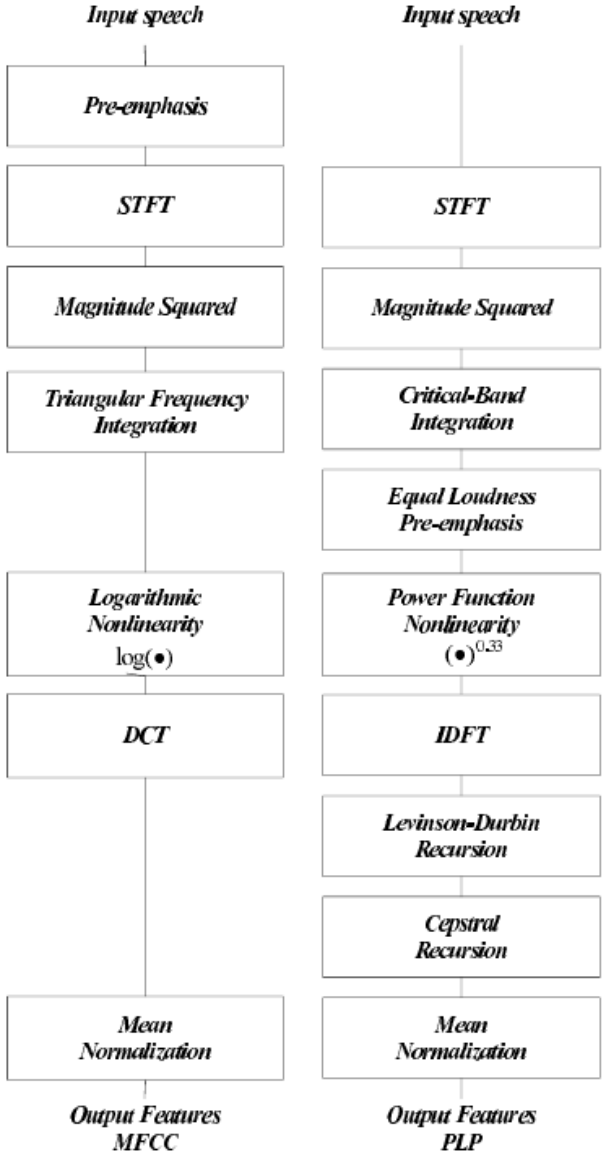

Fig. 4. Block diagrams of MFCC and PLP processing

Short-time Fourier Transform (STFT) analysis is performed using a hamming window, and triangular frequency integration is performed for spectral analysis. The logarithmic nonlinearity stage follows, and the final features are obtained through the us of a Discrete Cosine Transform (DCT) [5].

PLP processing, which is similar to MFCC processing in some ways, begins with STFT analysis followed by critical-band integration using trapezoidal frequency-weighting functions. In contrast to MFCC, pre-emphasis is performed based on an equal-loudness curve after frequency integration. The nonlinearity in PLP is based on the power-law nonlinearity proposed by Stevens. After this stage, Inverse Fast Fourier Transform (IFFT) and Linear Prediction (LP) analysis are performed in sequence. Cepstral recursion is also usually performed to obtain the final features from the LP coefficients.

The simplest way of performing normalization is using CMN or MVN. Histogram normalization (HN) is a generalization of these approaches. CMN is the most basic form of noise compensation schemes, and it can remove the effects of linear filtering if the impulse response of the filter is shorter then the window length [6].

By assuming that the mean of each element of the feature vector from all utterances is the same, CMN is also helpful for additive noise as well. CMN can be expressed mathematically as follows:

$$\overline{c}_i = c_i[j] - \mu_{c_i}, \quad 0 \le i \le I-1, \quad 0 \le j \le J-1,$$

where μ_{c_i} is the mean of the i^{th} element of the cepstral vector. In the above equation, $c_i[j]$ and $\overline{c}_i[j]$ represent the original and normalized cepstral coefficients for the i^{th} element of the vector at the j^{th} frame index. I denotes the dimensionality of the feature vector and J denotes the number of frames in the utterance.

MVN is a natural extension of CMN and is defined by the following equation:

$$\overline{c}_i[j] = \frac{c_i[j] - \mu_{c_i}}{\sigma_{c_i}}, \quad 0 \le i \le I-1, \quad 0 \le j \le J-1$$

where μ_{c_i} and σ_{c_i} are the mean and standard deviation of the i^{th} element of the cepstral vector [7].

Results

Fig. 5 compares the speech recognition accuracy obtained under various types of noisy conditions. We used subsets of 1600 utterances for training and 600 utterances for testing from the DARPA Resource Management 1 Corpus (RM1). In other experiments, which are shown in Fig. 5, we used the DARPA Wall Street Journal WSJ0-si84 training set and WSJ0 5k test set. For training the acoustical models we used SphinxTrain 1.0 and for decoding, we used Sphinx 3.8. For MFCC processing, we used sphinxe fe included in sphinxbase 0.4.1.

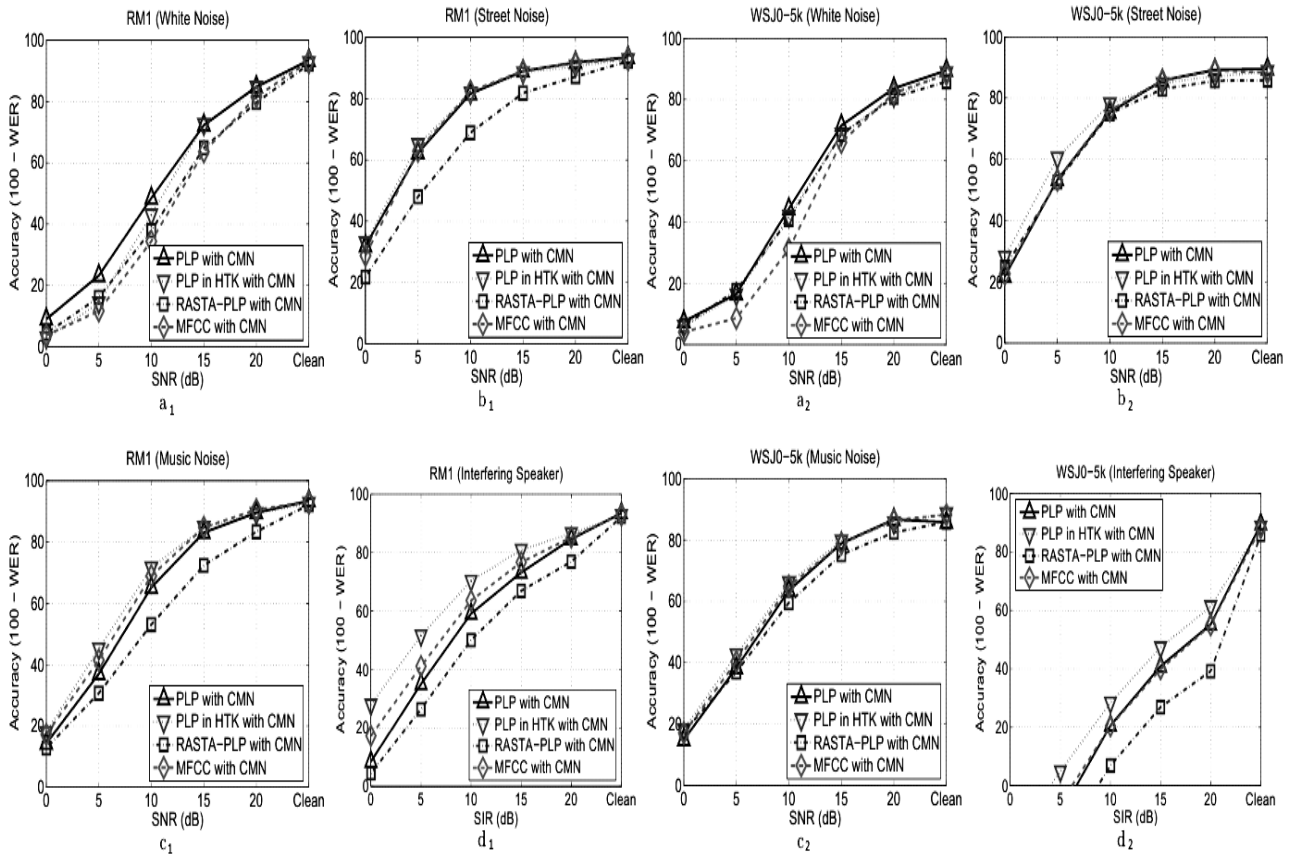

Fig. 5. Comparison of MFCC and PLP processing in different environments using the RM1 (WSJ0 5k) test set:
a_1, a_2 — additive white gaussian noise; b_1, b_2 — street noise; c_1, c_2 — background music;
d_1, d_2 — interfering speaker

For PLP processing, we used both HTK 3.4 and the MATLAB. Both of the PLP packages show similar performance, except for the for reverberation and interfering speaker environments, where the version of PLP included in HTK provided better performance. In all these experiments, we used 12^{th}-order feature vectors including the zeroth coefficient, along with the corresponding delta and delta-delta cepstra.

As shown in these figures, MFCC and PLP show provide speech recognition accuracy. Nevertheless, in our experiments we found that RASTA processing is not as helpful as conventional Cepstral Mean Normalization (CMN).

Conclusions

In the work described in later chapters of this thesis, we will develop an algorithm that is motivated by auditory observations, that imposes a smaller computational burden, and that can be implemented as an online algorithm that operates in sub-real time with only a very small delay. Instead of trying to estimate the environment function and maximizing the likelihood, which is very computationally costly, we will simply use the rate of power change or power distribution of the test utterance.

While it is generally agreed that a window length between 20 ms and 30 ms is appropriate for speech analysis, there is no guarantee that this window length will remain optimal for the estimation of or the compensation for additive-noise components. Since the noise characteristics are usually stationary compared to speech, it is expected that longer windows might be more helpful for noise compensation purposes. We note that even though longer duration windows may be used for noise compensation, we still need short duration windows for the actual speech recognition.

The Frequency scales describe how the physical frequency of an incoming signal is related to the representation of that frequency by the human auditory system. In general, the peripheral auditory system can be modeled as a bank of bandpass filters, of approximately constant bandwidth at low frequencies and of a bandwidth that increases in rough proportion to frequency at higher frequencies.

In addition to the nonlinearities we included another power-law nonlinearity which is an approximation to the physiological model of Heinz et al. between 0 and 50 dB SPL in the Minimum Mean Square Error (MMSE) sense. In this approximation, the estimated power coefficient is around 1/10.

PLP processing, which is similar to MFCC processing in some ways, begins with STFT analysis followed by critical-band integration using trapezoidal frequency-weighting functions. In contrast to MFCC, pre-emphasis is performed based on an equal-loudness curve after frequency integration.

We discussed several different rate-level nonlinearities based on different data.

Up until now, there has not been much discussion or analysis of the type of nonlinearity that is best for feature extraction.

For a nonlinearity to be appropriate, it should satisfy some of the following characteristics: it should be robust with respect to the presence of additive noise and reverberation; it should discriminate each phone reasonably well; the nonlinearity should be independent of the absolute input sound pressure level, or at worst, a simple normalization should be able to remove the effect of the input sound pressure level.

In other experiments, which are shown in Fig. 5, we used the DARPA Wall Street Journal WSJ0-si84 training set and WSJ0 5k test set.

For training the acoustical models we used SphinxTrain 1.0 and for decoding, we used Sphinx 3.8.

For MFCC processing, we used sphinxe fe included in sphinxbase 0.4.1.

For PLP processing, we used both HTK 3.4 and the MATLAB.

Both of the PLP packages show similar performance, except for the for reverberation and interfering speaker environments, where the version of PLP included in HTK provided better performance.

In all these experiments, we used 12^{th}-order feature vectors including the zeroth coefficient, along with the corresponding delta and delta-delta cepstra.

As shown in these figures, MFCC and PLP show provide speech recognition accuracy.

Nevertheless, in our experiments we found that RASTA processing is not as helpful as conventional Cepstral Mean Normalization (CMN).

REFERENCES

1. ***Lavrynenko O.*** Method of Voice Control Functions of the UAV / O. Lavrynenko, G. Konakhovych, D. Bakhtiiarov // Methods and Systems of Navigation and Motion Control (MSNMC), IEEE 4th International Conference. — 2016. — С. 47–50.

2. ***Bakhtiiarov D.*** Protected System of Radio Control of Unmanned Aerial Vehicle / D. Bakhtiiarov, G. Konakhovych, O. Lavrynenko // Methods and Systems of Navigation and Motion Control (MSNMC), IEEE 4th International Conference. — 2016. — С. 196–199.

3. A digital speech signal compression algorithm based on wavelet transform / G. F. Konakhovych, O. Y. Lavrynenko, D. I. Bakhtiiarov, V. V. Antonov // Electronics and control systems. — 2016. — №2. — С. 30–36.

4. Алгоритм сжатия сигналов речевых команд управления функциями беспилотного летательного аппарата / А. Ю. Лавриненко, Г. Ф. Конахович, Р. С. Одарченко, Д. И. Бахтияров // Авиационно-космическая техника и технология. — 2016. — №3. — С. 57–67.

5. Порівняльний аналіз перетворення Фур'є, косинусного перетворення та вейвлет-перетворення як спектрального аналізу цифрових мовних сигналів / Г. Ф. Конахович, О. І. Давлет'янц, О. Ю. Лавриненко, Д. І. Бахтіяров // Наукоємні технології. — 2015. — №3. — С. 210–220.

6. **Конахович Г.Ф.** Комп'ютерне моделювання захищеного каналу керування безпілотним літальним апаратом / Г. Ф. Конахович, Д. И. Бахтияров, А. Ю. Лавриненко // Наукоємні технології. — 2015. — №4. — С. 283–290.

7. Problems of unauthorized interference to the work of UAV and methods of its solving / I. O. Kozliuk, D. I. Bakhtiiarov, O. Y. Lavrynenko, I. V. Tretiak // Science-Based Technologies. — 2016. — №2. — С. 206–211.

8. **Lavrynenko O. Yu.** Compression algorithm of voice control commands of UAV based on wavelet transform / O. Yu. Lavrynenko, G. F. Konakhovych, D. I. Bakhtiiarov // Electronics and control systems. — 2018. — №1. — С. 17–22.

Lavrynenko O., Konakhovych G., Bakhtiiarov D.

COMPARATIVE ANALYSIS OF SPEECH RECOGNITION ALGORITHMS IN UAV VOICE CONTROL SYSTEM

The article proposes to perform a comparative analysis of the presented algorithms for processing voice control signals for an unmanned aerial vehicle, which can be implemented on processors with low computing power using online processing in real time. It is shown that these approaches are effective in improving the accuracy of speech recognition in the presence of various types of noise and a sound-reflecting control environment, which is an important problem in voice control systems for an unmanned aerial vehicle. An algorithm for calculating the mel-frequency cepstral coefficients, which appear in the role of the main features of speech recognition, is presented. A comparative analysis of two methods of distinguishing informative features of speech recognition in the voice control system of an unmanned aerial vehicle was made, namely, mel-frequency cepstral factors and the coefficients obtained with the aid of a linear prediction algorithm, where as a result of the conducted scientific experiment, under the influence of given

noise, it was concluded that in these problems, the optimal method of exclusion is the mel-frequency cepstral factors, since they show the best value for obsalutnomu criterion of speech recognition quality. The expediency of using the proposed system for recognizing voice commands of an unmanned aerial vehicle based on the cepstral analysis is substantiated and experimentally proved. The obtained results of the experimental research allow to draw a conclusion about the advisability of further practical application of the developed system for recognizing voice commands for the control of an unmanned aerial vehicle on the basis of a cepstral analysis.

Keywords: MEL scale; BARK scale; UAV; speech recognition; MFCC; minimum mean square error.

Лавриненко О. Ю., Конахович Г. Ф., Бахтіяров Д. І.

ПОРІВНЯЛЬНИЙ АНАЛІЗ АЛГОРИТМІВ РОЗПІЗНАВАННЯ МОВИ В СИСТЕМІ ГОЛОСОВОГО УПРАВЛІННЯ БПЛА

У статті пропонується провести порівняльний аналіз представлених алгоритмів обробки сигналів голосового управління безпілотним літальним апаратом, які можуть бути реалізовані на процесорах з малою обчислювальною здатністю використовуючи онлайн-обробку в режимі реального часу. Показано, що запропоновані підходи ефективні в поліпшенні точності розпізнавання мови при наявності різних типів шумів і звуковідбиваючого середовища управління, що є важливою проблемою в системах голосового управління безпілотним літальним апаратом. Представлений алгоритм обчислення мел-частотних кепстральних коефіцієнтів, які виступають в ролі основних ознак розпізнавання мови. Був проведений порівняльний аналіз двох методів виділення інформативних ознак розпізнавання мови в системі голосового управління безпілотним літальним апаратом, а саме мел-частотні кепстральні коефіцієнти і коефіцієнти отримані за допомогою алгоритму лінійного передбачення, де в результаті проведеного наукового експерименту при впливі заданих шумів були зроблені висновки, що в даних задачах оптимальним методом виділення є мел-частотні кепстральні коефіцієнти, так як вони показують найкращий показник по абсолютному критерію якості розпізнавання мови. Обґрунтовано та експериментально доведено доцільність використання запропонованої системи розпізнавання голосових команд управління безпілотним літальним апаратом на основі кепстрального аналізу. Отримані результати експериментального дослідження дозволяють зробити висновок про доцільність подальшого практичного застосування розробленої системи розпізнавання голосових команд управління безпілотним літальним апаратом на основі кепстрального аналізу.

Ключові слова: шкала MEL; шкала BARK; БПЛА; розпізнавання мови; MFCC; мінімальна середньоквадратична похибка.

Лавриненко А. Ю., Конахович Г. Ф., Бахтияров Д. И.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ РАСПОЗНАВАНИЯ РЕЧИ В СИСТЕМЕ ГОЛОСОВОГО УПРАВЛЕНИЯ БПЛА

В статье предлагается произвести сравнительный анализ представленных алгоритмов обработки сигналов голосового управления беспилотным летательным аппаратом, которые могут быть реализованы на процессорах с малой вычислительной способностью используя онлайн-обработку в режиме реального времени. Показано, что эти подходы эффективны в улучшении точности распознавания речи при наличии различных типов шумов и звукоотражающей среды управления, что является важной проблемой в системах голосового управления беспилотным летательным аппаратом. Представлен алгоритм вычисления мел-частотных кепстральных коэффициентов, которые выступают в роли основных признаков распознавания речи. Был проведен сравнительный анализ двух методов выделения информативных признаков распознавания речи в системе голосового управления беспилотным летательным аппаратом, а именно мел-частотные кепстральные коефициенты и коефициенты, полученные с помощью алгоритма линейного предсказания, где в результате проведенного научного експеременте при воздействии заданных шумов были сделаны выводы, что в данных задачах оптимальным методом выдиления является мел-частотные кепстральные коефициенты, так как они показывают наилучший показатель по обсалютному критерию качества распознавания речи. Обосновано и экспериментально доказано целесообразность использования предложенной системы распознавания голосовых команд управления беспилотным летательным аппаратом на основе кепстрального анализа. Полученные результаты экспериментального исследования позволяют сделать вывод о целесообразности дальнейшего практического применения разработанной системы распознавания голосовых команд управления беспилотным летательным аппаратом на основе кепстрального анализа.

Ключевые слова: шкала MEL; шкала BARK; БПЛА; распознавание речи; MFCC; минимальная среднеквадратическая ошибка.

Стаття надійшла до редакції 01.04.2018 р.
Прийнято до друку 04.06.2018 р.
Рецензент — д-р техн. наук, проф. Сібрук Л. В.