Pobrane z czasopisma Annales H - Oeconomia **http://oeconomia.annales.umcs.pl**
Data: 18/05/2026 19:37:43

DOI:10.17951/h.2018.52.2.71-78

ANNALES
UNIVERSITATIS MARIAE CURIE-SKŁODOWSKA
LUBLIN – POLONIA

VOL. LII, 2 SECTIO H 2018

* University of Warsaw. Faculty of Economic Sciences
** Profeosoft, Kraków

*KAROLINA KULIGOWSKA,
**PAWEŁ KISIELEWICZ, **ALEKSANDRA WŁODARZ

kkuligowska@wne.uw.edu.pl, profeosoft@gmail.com, aleksandra.rojek@profeosoft.pl

Managing Development of Speech Recognition Systems: Performance Issues

Zarządzanie rozwojem systemów rozpoznawania mowy: problemy wydajności

Keywords: speech recognition system; speech-to-text performance; STT development

Słowa kluczowe: system rozpoznawania mowy; wydajność rozpoznawania mowy; rozwój rozpoznawania mowy

JEL Code: D83; D89; L86

Introduction

Designing a machine that mimics the human ability to listen and respond appropriately to the spoken language has fascinated engineers and scientists for centuries. The realization of this dream began with children's toy Radio Rex produced in the 1920s. A celluloid dog responded to its name by popping out of its cardboard house. The dog moved when spring was released as a result of the appropriate dose of acoustic energy in the range of 500–700 Hz. The first formant vowel "e" has the approximate frequency of 500 Hz, which can be triggered by speaking the word "rex" [Morgan 2012, p. 7].

Since late the 1970s the present day researches on speech recognition have achieved significant progress [Anusuya, Katti, 2009, pp. 186–191; Virtanen, Singh, Raj (eds.), 2013, pp. 9–30]. Currently built systems react to spontaneous speech and

Pobrane z czasopisma Annales H - Oeconomia **http://oeconomia.annales.umcs.pl**
Data: 18/05/2026 19:37:43

take into account the statistical variability of natural language. However, the quality of automatic speech recognition is affected by many factors, therefore so much effort is put into managing development of the performance of speech recognition systems.

The aim of this paper is to examine performance issues in speech recognition systems and indicate the direction of managing development and improvement of these systems. As a preliminary work, we reviewed over 100 recent papers that contain the most important results of research in the design, construction, and implementation of speech recognition systems. In reference to this paper, we included publications which proved to be the most significant for our diagnosis of performance issues of speech recognition systems.

The paper is organized as follows. Section 1 presents design and functioning of existing speech recognition systems. Section 2 analyzes drawbacks and limitations of speech recognition systems. Section 3 presents our conclusions.

1. Design and functioning of speech recognition systems

Speech recognition systems base their architecture on similar modules. These modules are responsible for processing the input data according to certain rules. In the first step, the speech signal is subjected to pretreatment process, when a selection and extraction of the characteristics of the signal are made. In the next step, the classification of these characteristics is carried out using different methods of matching patterns.

Initially, the microphone takes a sample of sound which is converted to the digital signal. It enables the speech analysis through the computer. Due to the fact that audio data has a lot of information such as additive noises, which arise during sampling and humming, this signal is subjected to the pretreatment processing. Thus, this phase reduces the amount of unnecessary information (i.e. compression and noise reduction).

The key tasks of speech recognition are the segmentation and feature extraction. Before the main recognition process, the speech signal usually must be divided into small segments. It can be divided into words, phonemes, syllables, but also into parts depending on the phonetic features. Very important in speech segmentation is also to separate the speech from silence [Ziółko, Ziółko, 2011, pp. 224–225]. Feature extraction constitutes a conversion of a speech signal into feature vectors. This representation is used for further analysis and processing. It plays a very important role as it distinguishes one word from another, which largely affects system performance. The main speech signals features are amplitude, power, intensity, fundamental frequency [Shanthi, Chelpa, 2013, pp. 481–483].

Following the feature extraction, the next stage – which is also called decoding – is based on recognizing separated speech signals. The decoder, that is the engine responsible for finding the best fit in the knowledge base based on the incoming feature vectors, constitutes a key component of speech recognition system [Gaikwad,

Pobrane z czasopisma Annales H - Oeconomia **http://oeconomia.annales.umcs.pl**
Data: 18/05/2026 19:37:43

Gawali, Yannawar, 2010, pp. 18–22; Gubka, Kuba, Jarina, 2013, pp. 565–569]. It matches a detected word to a word already known from the gathered knowledge base using whole-word matching or sub-word matching.

2. Drawbacks and limitations of speech recognition systems

The performance of speech recognition systems is influenced by many factors. We analyzed them and distinguished 11 main areas, namely: recognition environment, recording equipment, prosody, accent, children's speech, emotional state, adjustment to the speaker's voice, limited vocabulary, homonyms, context, Slavic languages. We describe them in following subsections.

2.1. Recognition environment

The environment in which system is running is very important. If during the sampling unwanted noises in the background are recorded, accurate recognition cannot be provided. Noise causes two main effects on the speech representation: introduces distortions in the representation of space, and, due to their random nature, causes loss of information. Distorting representation of space causes discrepancies between the training conditions (clean environment) and conditions under which recognition takes place (noisy environment) [Akbarinia, Valdez Medrano, Zamani, 2011, p. 5; Wu, Liu, 2012, p. 1]. The loss of information caused by the noise significantly reduces efficiency and understanding, even with optimum compensation of discrepancies.

2.2. Recording equipment

The problem closely associated with the noise and recognition environment is the quality of a device that records the voice. If the microphone is not sensitive enough or is too sensitive, it can create audio information that will be difficult to decipher. This is especially important when the microphone is so sensitive that speech becomes distorted, which makes software almost useless.

2.3. Prosody

Prosody plays an important role in understanding spoken language: it helps to recognize spoken words, deal with ambiguities and with the discourse structure [Mary, 2012, p. 17; Seppi, Demuynck, Compernolle, 2011, p. 1]. Lexical segmentation component in the processing of human speech is based on terms, which in turn are based on prosody. Prosody also helps in locating boundaries between words. In many languages, we can distinguish words based only on the internal prosodic structure. However, prosody is difficult to model, and prosodic features are highly

Pobrane z czasopisma Annales H - Oeconomia **http://oeconomia.annales.umcs.pl**
Data: 18/05/2026 19:37:43

variable [Anumanchipalli, Oliveira, Black, 2012, p. 153]. Given the same sentence, speakers have freedom to focus on any concept they choose, and the degree to which the emphasis is laid on a given word.

2.4. Accent

Different people have their own style of speaking, depending on many factors, such us dialect and accent, socioeconomic background and contextual variables such us the degree of familiarity between the speaker and hearer. The individual differences often cause many difficulties in modeling the speaker-independent systems intended for processing the output data from any variant of the language. The past few decades have seen considerable progress in automatically identifying language. Accent and dialect recognition have begun to receive attention from the field of speech technology. The task of dialect identification is to recognize speaker's regional dialect, within a predetermined language [Biadsy, 2011, pp. 1–2].

2.5. Children's speech

Children's speech is still poorly understood area in the field of computer speech recognition, and the problem is mainly related with physical aspects. Children compared to adults have shorter vocal tract and vocal folds. This results in a higher fundamental frequency, which is reflected in a large distance between the harmonic, resulting in poor spectral resolution of voiced sounds. Another problem is the incorrect pronunciation of children, as very often they do not know how to articulate specific phonemes. Moreover, children's vocabulary is usually very small, it can include words that do not appear in grown-ups speech. What is more, very often they do not know correct inflectional forms of certain forms, especially those words that are exceptions to common rules. Although applying several techniques designed to improve the accuracy of children's speech recognition systems, efficiency of such systems is still much lower than in the case of an adult speech recognition.

2.6. Emotional state

Emotional state greatly affects the speech spectrum. It is known, that changes in speakers mood have a substantial impact on the features extracted from his speech, hence directly affect the accuracy of speech recognition.

2.7. Adjustment to the speaker's voice

Large variation of the voice acoustic realization can be reduced by adapting the system to the user. Speaker-dependent systems provide better performance and accuracy of recognition than speaker independent system, whereby it should be

Pobrane z czasopisma Annales H - Oeconomia **http://oeconomia.annales.umcs.pl**
Data: 18/05/2026 19:37:43

noted that systems which adapt to the user are not applicable in all situations, and often are less practical [Cloarec, Jouvet, 2008, p. 4529]. Nevertheless, such systems require sufficient time to learn the speaker's voice and the way of speaking. Even if the system has been thoroughly trained, it still makes mistakes. What is more, the human voice is also variable by periodic changes in voice, caused by cold, hoarseness, stress or different emotional states, and changes in the voice during puberty.

2.8. Limited vocabulary

Most speech recognition systems have limited vocabulary. This involves problems with recognizing unusual or fictitious names and words that are not in the vocabulary. Out of vocabulary words (OOVW) are unknown words that appear during speech testing, but not in the recognition dictionary [Qin, 2013, pp. 13–17]. Typically, these words carry with them relevant content such as names and locations which contain information crucial to the success of many speech recognition tasks. However, it is not possible to define the lexicon, which would contain all the words that can occur during speech recognition.

2.9. Homonyms

Homonyms are words that sound the same but have different meanings. Speech recognition systems are not able to distinguish them based only on sound, and very often they are difficult to distinguish even by a human. When the system has a choice of several identical variants, it will not always make the right selection. Some systems try to cope with this problem by analyzing the context using statistical models that select the most likely words which are translated. However, these solutions do not work well for recognizing isolated words, where there is no context at all.

2.10. Context

One of the most complex problems of speech recognition is to determine the context in which words were spoken. Computer software is unable to identify the intended meaning of the words which leads to several problems. Some words that sound very similar may be correctly recognized only when their context is known.

2.11. Slavic languages

Most of the speech recognition system operate on the most widely spoken languages, such as English, German, French or Japanese. Whereas there are languages which still are waiting for the rapid development of modern speech technologies. Especially the group of Slavic languages constitutes a great challenge for researchers – mainly because of the inflectional nature of Slavic languages [Janicki, Wawer, 2011, p. 713].

Pobrane z czasopisma Annales H - Oeconomia **http://oeconomia.annales.umcs.pl**
Data: 18/05/2026 19:37:43

All Slavic languages exhibit very large degree of inflection. The vast majority of lexical items modify its basic form according to grammatical, morphological and contextual relations. Morphology of Slavic languages is even more complex. New words can be created by adding (single or multiple) prefixes, suffixes, and endings to a stem, or also by modifying the stem itself. All these specific characteristics give rise to many forms of expression. This difference is very large in comparison with speech recognition systems designed e.g. for English, in which the dictionary for 50,000 most commonly used words gives a coverage ratio of 99%. Slavic languages generally require dictionaries, which are 10 to 20 times greater [Nouza et al., 2010, pp. 227–229].

Authors of this paper are all Polish, therefore it is natural that our primary interest goes to Slavic languages. However, due to small amount of work concerning difficulties of Slavic language inflection and free order syntax, issues discussed above concern also other, little widespread, non-Slavic languages.

Conclusions

One of the fundamental challenges for current research on speech technology in understanding and modeling performance issues that need to be managed in creating speech recognition systems. We identified 11 sources of performance issues of speech recognition systems that require managing development: recognition environment, recording equipment, prosody, accent, children's speech, emotional state, adjustment to the speaker's voice, limited vocabulary, homonyms, context, Slavic languages. These sources of variability we can group into three main areas, such as recognition environment, speaking style, and speaker characteristics.

Recognition environment in which the speech is recorded has a big impact on the performance of recognition. Suitable conditions for recording speech and well-chosen hardware configuration are principal to a good acoustic signal recognition. Therefore speech recognition software often requires that the words are spoken clearly without any additional noise that can come from many sources. The user should work in a quiet location, using a good quality microphone, which should be placed close to the mouth.

Speaking style constitutes another factor that affects the quality of speech recognition. Isolated word systems require to leave short pauses between words. Due to the fact that this is an unnatural speaking style, those systems lose their popularity in favor of continuous speech recognition systems, on which most research is focused.

Speaker characteristics are another source of performance issues in speech recognition systems. Every person has different personal features (health condition, emotional state), different way of speaking, different tone and timbre, speaks at different rates and rhythm, pronounces words differently, and uses different language. Further differences arise from demographic, cultural and geographic alternations, such as age, sex, class affiliation, and accent. The solution for such

Pobrane z czasopisma Annales H - Oeconomia **http://oeconomia.annales.umcs.pl**
Data: 18/05/2026 19:37:43

variability is to construct a speaker-dependent speech recognition system, which will make fewer errors. However, this is related to the construction of a separate system for each speaker, which, taking into account all the people living on Earth, would be unachievable.

References

Akbarinia A., Valdez Medrano J., Zamani R., *Speech Recognition for Noisy Environments – Feasibility of Voice Command in Construction Settings*, Engineer's thesis, Department of Computer Science and Engineering, University of Gothenburg, Goteborg, 2011.

Anumanchipalli G.K., Oliveira L.C., Black A.W., *Intent transfer in speech-to-speech machine translation*, "IEEE Workshop on Spoken Language Technology" 2012, **DOI: https://doi.org/10.1109/SLT.2012.6424214**.

Anusuya M.A., Katti S.K., *Speech Recognition by Machine: A Review*, "International Journal of Computer Science and Information Security" 2009, Vol. 6(3).

Biadsy F., *Automatic Dialect and Accent Recognition and its Application to Speech Recognition*, Department of Computer Science, Columbia University, 2011 (doctoral dissertation).

Cloarec G., Jouvet D., *Modeling inter-speaker variability in speech recognition*, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2008, **DOI: https://doi.org/10.1109/ICASSP.2008.4518663**.

Gaikwad S., Gawali B., Yannawar P., *A review on Speech Recognition Technique*, "International Journal of Computer Applications" 2010, Vol. 10(3), **DOI: https://doi.org/10.5120/1462-1976**.

Gubka R., Kuba M., Jarina R., *Universal approach for sequential audio pattern search*, Proceedings of the 2013 Federated Conference on Computer Science and Information Systems FedCSIS, Annals of Computer Science and Information Systems, Kraków 2013.

Janicki A., Wawer D., *Automatic Speech Recognition for Polish in a Computer Game Interface*, Proceedings of the 2011 Federated Conference on Computer Science and Information Systems FedCSIS, Annals of Computer Science and Information Systems, Szczecin 2011.

Mary L., *Extraction and Representation of Prosody for Speaker*, Speech and Language Recognition, SpringerBriefs in Electrical and Computer Engineering, Springer, New York 2012, **DOI: https://doi.org/10.1007/978-1-4614-1159-8**.

Morgan N., *Deep and Wide: Multiple Layers in Automatic Speech Recognition*, "IEEE Transactions on Audio, Speech and Language Processing" 2012, Vol. 20(1), **DOI: https://doi.org/10.1109/TASL.2011.2116010**.

Nouza J., Zdansky J., Cerva P., Silovsky J., *Challenges in Speech Processing of Slavic Languages (Case Studies in Speech Recognition of Czech and Slovak)*, [in:] A. Esposito, N. Campbell, C. Vogel, A. Hussain, A. Nijholt (eds.), *Development of Multimodal Interfaces: Active Listening and Synchrony*, Springer Verlag, Berlin–Heidelberg 2010.

Qin L., *Learning Out-of-Vocabulary Words in Automatic Speech recognition*, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh 2013 (doctoral dissertation).

Seppi D., Demuynck K., Compernolle D. van, *Template-based Automatic Speech Recognition meets prosody*, 12th Annual Conference of the International Speech Communication Association (Interspeech 2011), Florence 2011.

Shanthi T., Chelpa L., *Review of Feature Extraction Techniques in Automatic Speech Recognition*, "International Journal of Scientific Engineering and Technology" 2013, Vol. 2(6).

Virtanen T., Singh R., Raj B. (eds.), *Techniques for Noise Robustness in Automatic Speech Recognition*, Wiley, London 2013.

Pobrane z czasopisma Annales H - Oeconomia **http://oeconomia.annales.umcs.pl**
Data: 18/05/2026 19:37:43

Wu C.-H., Liu C.-H., *Robust Speech Recognition for Adverse Environments*, [in:] S. Ramakrishnan (ed.), *Modern Speech Recognition Approaches with Case Studies*, Intech 2012, **DOI: https://doi.org/10.5772/47843**.
Ziółko B., Ziółko M., *Przetwarzanie mowy*, Wydawnictwa AGH, Kraków 2011.

Zarządzanie rozwojem systemów rozpoznawania mowy: problemy wydajności

Rozpoznawanie mowy umożliwia przekształcanie wypowiadanych słów i zdań w tekst w formie cyfrowej. Technologia ta jest od wielu lat przedmiotem licznych badań naukowych oraz komercyjnych. Celem niniejszego artykułu jest zbadanie zagadnień dotyczących wydajności systemów rozpoznawania mowy i zarządzanie rozwojem tych systemów. Dogłębna analiza w zakresie ograniczeń wydajnościowych systemów rozpoznawania mowy pozwoliła na zidentyfikowanie problemów, które trzeba przezwyciężyć. Wskazują one kierunek zmian w zarządzaniu rozwojem systemów rozpoznawania mowy.

Managing Development of Speech Recognition Systems: Performance Issues

Speech recognition enables the transformation of spoken words and sentences into text in digital form. This technology is a subject of numerous studies and commercial development for many years. The aim of this paper is to examine performance issues of speech recognition and to manage the development in this field. Thorough analysis of performance limitations of speech recognition systems we identified main 11 issues to overcome. They indicate the direction of managing development of speech recognition systems.