CMST 27(2) 41–55 (2021) DOI:10.12921/cmst.2021.0000015

Automatic Speech Recognition and its Application to Media Monitoring

R. Cecko, J. Jamroży, W. Jęśko, E. Kuśmierek*, M. Lange, M. Owsianny

Poznan Supercomputing and Networking Center
ul. Jana Pawła II 10
61-139 Poznan, Poland
**E-mail: ewa.kusmierek@man.poznan.pl*

Received: 21 May 2021; revised: 22 June 2021; accepted: 24 June 2021; published online: 29 June 2021

Abstract: In this paper we present application of the automatic speech recognition technology in the area of media monitoring. We describe the use of computational models and methods by two ASR technologies, namely a Hidden Markov Model with a Gaussian Mixture Model and Deep Neural Networks, that were crucial in the ASR development. Both approaches were implemented in our speech recognition ARM-1 engine developed for the Polish language. We provide details on the implementation choices, specifically adjustments made for media monitoring application guided by the characteristics of media content. Performance of both versions of our engine is evaluated and compared.

Key words: signal processing, automatic speech recognition, machine learning, neural networks, media monitoring

I. INTRODUCTION

Automatic speech recognition (ASR) methods have been under development for several decades now. The initial success was achieved with an application of statistical modeling with Hidden Markov Models (HMM). In this approach machine learning algorithms are used to build models based on training data in order to process and identify the speech signal. HMMs combine various types of information such as acoustics, language and syntax in one probabilistic system.

The speech signal is a result of many mental and physical processes and as such is characterized by large variability which has intrapersonal and interpersonal sources. Therefore, a large training set is necessary to build proper statistical models. A model dedicated to a certain type of speech, e.g. speech of a certain age group, can often capture the nature of the speech signal better than a more general model. Typically an area of ASR application determines the type of speech to be recognized and its acoustic properties.

The HMM-based approach dominated till late 1990s when it became obvious that it is not sufficient for more challenging cases such as processing spontaneous speech recorded in a noisy environment with many speakers' voices overlapping or for processing an audio signal which is inaudible for a human. Deep learning methods are gradually taking over various aspects of the speech recognition process.

In this paper we present ARM-1 engine [1, 2] developed for speech recognition of the Polish language and its application in one area that can be particularly demanding, namely media monitoring. We are faced with rapid development of information technology. Information can be easily created and distributed, which results in a constantly growing amount of information available in the media and increased dynamics of its spread. Therefore, automation of media content processing and analysis performed in various media monitoring areas is necessary to deal with the sheer volume of content and its diversity. Another area that could benefit from automatic content analysis is digital humanities [3] as a discipline relying on advanced computational methods and experiencing dynamic growth with increasing availability of various tools and services.

As already mentioned, ASR task becomes more challenging when we deal with spontaneous speech recorded

under less than ideal acoustic conditions and with several people talking at the same time. This characteristic is very common for audio content presented in the media. On top of that audio may be characterized by a high degree of variability due to frequent speaker and acoustic conditions changes, making it difficult for an ASR system to adapt. The HMM-based approach used in the first version of ARM-1 engine did not meet this challenge. Application of deep neural networks (DNN) was needed for media monitoring to fully benefit from speech recognition automation.

In this paper we provide a brief introduction to the ASR technology and the use of computational models and methods in this area. We describe both types of methods used in automatic speech recognition, namely HMM and DNN, and explain how they were implemented in the subsequent versions of the ARM-1 engine. By engine we mean software which implements automatic speech recognition functionality and which is the heart of a number of tools and systems developed for various ASR applications.

The paper is organized as follows. In Section II we briefly describe related work on ASR applications to media monitoring. Section III provides background information on HMM and DNN-based approaches to ASR and describes which methods were implemented in the ARM-1 engine. We also provide information on the adjustments that were needed to successfully use the ARM-1 engine in media monitoring. Section IV describes test results obtained for two subsequent versions of ARM-1 engine regarding speech recognition accuracy and their analysis. We conclude the paper with the summary and conclusions in Section V.

II. RELATED WORK

A number of solution have been reported regarding the use of machine learning for automatic transcription of audio and video files in the media monitoring area. The vast majority of them present various dedicated speech recognition models for converting speech to text from recordings originating from one particular source, which is broadcast news [4–15]. Despite the ability to extract a lot of information from this source, it is important to note the low versatility of automatic transcription solutions that are dedicated only to one source. ASR systems dedicated to a specific source (in this case, broadcast news), and thus trained only using files from this source, are characterized by low versatility and applicability to other audio-video sources.

Recordings from broadcast news are characterized by good quality source files and, most importantly, the absence of external noise that can significantly affect the final speech recognition result. Learning ASR models on such ideal recordings prevents the effective use of these systems for recordings characterized by poorer quality and the possibility of additional ambient noise. The literature also describes a more universal ASR solution for media monitoring. The authors of this solution emphasize the possibility of using the realized ASR system in the field of automatic transcription of audio-video files from such sources as general and news television and radio channels [16].

The ASR solutions presented in the literature are based on a hybrid solution GMM/HMM [4–6, 14, 15, 17, 18], MLP-HMM [7–11, 16], DNN/HMM [17–19], time delay DNN (TDNN) [12, 18, 20], time delay DNN with projected LSTM (TDNN-LSTM) [18], acoustic model, pronunciation and language model based on the vocabulary of the language for which the system is dedicated, including English [9, 14, 16, 17], Portuguese [7–11, 16, 21], Spanish [9, 16], Greek [14], Lithuanian [12], Latvian [5], Slovenian [6], Ukrainian [13, 19] and Macedonian [15].

Research on both hybrid (acoustic model based on: GMM-HMM, DNN/HMM, TDNN, TDNN-LSTMP) and end-to-end solutions (LSTM) is presented in [18]. The author presented research on the end-to-end architecture proposed in [22], which consisted of a recurrent neural network (RNN) based on long short-term memory (LSTM) with adoption of connectionist temporal classification (CTC) objective function and decoding with the use of weighted finite-state transducers (WFST).

Accuracy of speech recognition of the media monitoring systems described in the literature is on various levels. The efficiency of the presented solutions is strictly related to a given language and the test set used in the evaluation process which was assembled for that language. The best results were achieved by systems based on hybrid solutions (mainly DNN-HMM and TDNN) with word error rate below 15%, i.e. 8.1% for Estonian [20], 9.3% for German [18], 11.9% for Ukrainian [13], 14.7% for Lithuanian [12], 14.9% for English [14]. In the case of systems supporting more than one language, the best results were achieved for the primary language for which a given system was built, e.g. in the case of the AUDIMUS.MEDIA [9, 11, 16], the following word error rates were reported: 18.4% for European Portuguese, 20.4% for English, 20.8% for Brazilian Portuguese and 21.2% for Spanish.

The ARM-1 engine presented in this paper is not dedicated to a specific media information source (broadcast news). Its main feature of the implemented ASR system is its high versatility and the possibility of using it with many different media information sources, including broadcast news, radio recordings, sports recordings and even different quality audio-video files available on the Internet (e.g. on the YouTube platform).

III. ASR TECHNOLOGIES AND THEIR IMPLEMENTATION IN ARM-1

In explaining methods used in automatic speech recognition it is crucial to understand the nature of speech signal that determines the difficulty of this process. Hence, in this Sec-

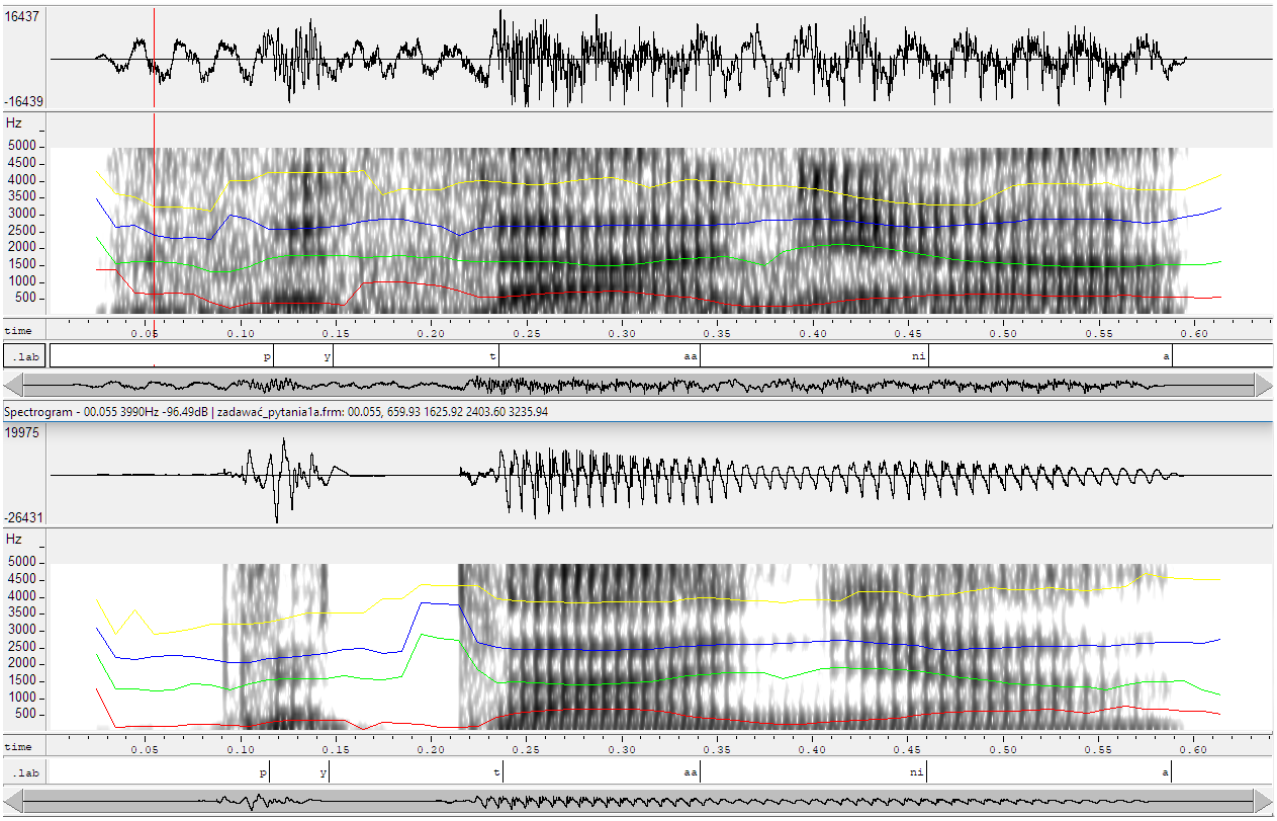


Fig. 1. Waveform and spectrogram of the word "pytania" recorded in the noisy environment and in a recording studio

tion we first describe the source and nature of speech variability. Next, we explain how this characteristic was dealt with in two main groups of ASR methods. For both, HMM and DNN-based methodology the description is followed by information on their implementation in the ARM-1 engine.

III. 1. Speech Signal and its Variability

The speech signal is a result of many mental and physical processes. The immanent feature of the speech sound is its variability and redundancy of the information it carries. The sources of variability can be divided into intrapersonal and interpersonal. A speaker, deciding on the content of speech, determines the manner of utterance (rate, loudness, accuracy of articulation, vocal effort, dictation, spontaneous speech). Realization of this decision also depends on the mental state, stress level, general health, social background of the speaker, as well as their upbringing, education, and the region of the country where they grew up. All of the aforementioned differences in voice are related to the within-person variability. The final effect in the form of a speech signal is also dependent on the anatomical structure of the vocal tract related to gender, age, past illnesses, i.e. inter-personal variability. These factors are compounded by the effect of auditory feedback. The articulatory organs are constantly monitored, corrected and compared with the speaker's intention so that the sound is clear and intelligible. Here we are dealing with the effects of coarticulation, i.e. similarity of adjacent sounds and many other psycholinguistic and acoustic effects.

The sources of variability are also other people's conversations, interference, noise, ambient noise and the acoustics of the room in which we perceive speech (reverberation, reflections). With too much noise the so-called Lombard effect can occur, i.e. an unintentional change in acoustic characteristics (tempo, duration of syllables, voice intensity, shift of the spectrum towards higher frequencies). Damaged hearing can interfere with the feedback. Before reaching the speech recognition system, the signal must be recorded by a microphone and sent through the electroacoustic track (preamplifier, amplifier, analog-to-digital converter) to the system, where it will be converted into a sequence of parameters – a feature vector. The electroacoustic track and its characteristics, quality and introduced distortions may cause an additional variability.

Fig. 1 shows the waveform and spectrogram of the Polish word "pytania" ("questions") obtained with the WaveSurfer software from a radio broadcast (noisy upper part) and the same word spoken naturally in the studio (lower part). The formants frequency contours are marked with colored lines. The transcription pane located in the figure below the spectrograms presents segmentation into individual phonemes of the utterance. In the noisy spectrogram we can

observe additional broadband ambient noise with a fairly energy-balanced frequency spectrum with a hum at the bottom of the band.

The speech recognition system must compensate for intrapersonal, interpersonal and hardware variability of the speech signal in order to make it independent of the speaker. A method widely used for this purpose turns out to be a statistical approach based on the construction of acoustic models created from large acoustic databases in which all of the above described effects occur [23].

III. 2. GMM/HMM-based Approach

The human auditory system can correctly recognize words, most of the time anyway, independent of who pronounces the word, the way it is pronounced and the context in which it is pronounced. In other words, we are able to compensate for various sources of variability in the speech signal. An ASR system must be able to do the same. One way to achieve this goal is to use machine learning methods to build a statistical model that captures characteristics necessary to distinguish and recognize various elements of the speech signal.

Given high variability of the speech signal, building such a model requires an enormous amount of data, namely recordings of utterances of words spoken in a given language. The acoustic model must be representative of speech that will be processed by the ASR system. Hence, building a universal model capable of representing utterances exhibiting variability from various sources is more challenging than building a model dedicated to some type of speech signal and its acoustic characteristics.

In building a statistical model we assume that the speech signal can be represented by a set of statistically independent phonetic-acoustic parameters, so called feature vector. Parameters are selected in such a way that their information volume is significantly smaller than the information volume of the source signal and at the same time they represent the source signal sufficiently well. The goal of the model training procedure is to build a recognition network consisting of word pronunciations defined as sequences of triphones. A triphone is a linguistic unit of sound called phonem with a certain left and right neighborhood which is used to capture the fact that pronunciation depends on the preceding and succeeding sound. Each triphone is modeled with a HMM consisting of a set of states as illustrated in Fig. 2. Each of the emitting states is described with sets (mixtures) of Gaussian probability densities (GMM) determined for each element of the feature vector. Connections between the states governed by the discrete probability a_{ij}, represent sequences of sounds that can occur. During the speech recognition process input data is divided into a number of observations. For each observation the value of each feature vector element is computed and then evaluated against GMMs $b_j(o_t)$ in order to obtain acoustic likelihood of the observation being a sound represented by a given HMM state.

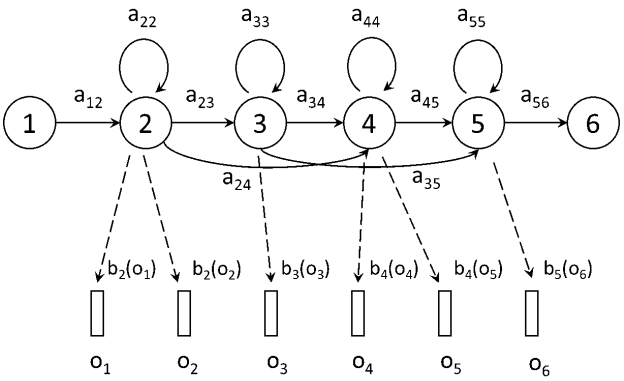


Fig. 2. Statistical HMM model of a phonem [24]

The recognition network is searched for the best path matching a sequence of input observations where path "goodness" is measured by cumulative likelihood.

Acoustic features are not the only elements used in determining the best hypothesis corresponding to the best path through the recognition network. The other one is linguistic probability that, roughly speaking, expresses probability of a word or phrase occurring in a given language. The linguistic model used to determine this probability is a statistical model generated by processing a large text corpus. The N-gram model captures probability that a word occurs provided that it is preceded (or succeeded) by a given sequence of $n-1$ words, where $n \leq N$ (order of the model). Similarly to the acoustic model, building a universal linguistic model requires an enormous amount of training data since such a model should capture a wide range of vocabulary and speech styles. Dedicated models can be built more easily for a specific subject area, for example related to judiciary vocabulary.

Unigram word probability (1-gram) is used in evaluating a path through a recognition network during speech processing in addition to the acoustic probability. In fact a number of higher probability hypotheses are selected and further evaluated using a higher order linguistic model in order to choose the best hypothesis. More precisely, a weighted average of the acoustic and linguistic probability is used for that evaluation.

III. 3. GMM/HMM-based ARM-1 Engine

The GMM/HMM speech recognition methodology has been implemented in the first version of the ARM-1 engine [25]. In this section we provide information on the engine implementation details followed by a description of the adjustments made in order to adapt the engine to the requirements imposed by its application to media monitoring.

III. 3. 1. ARM-1 Engine Architecture

The ARM-1 engine is made up of the following modules: DSP (Digital Signal Processing), Decoder and Rescorer

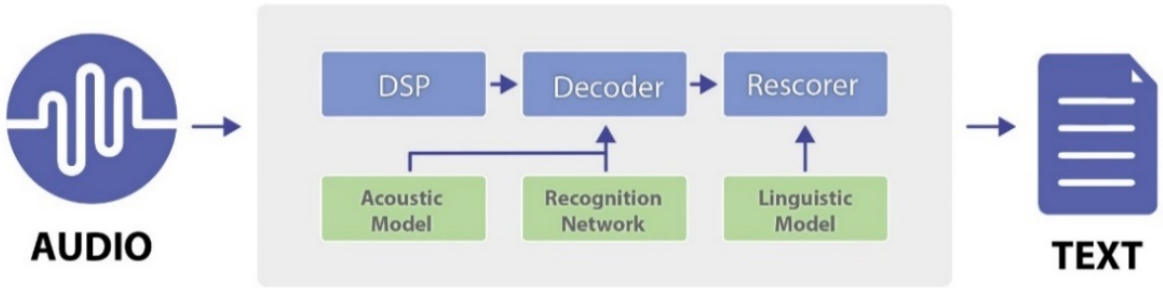


Fig. 3. ARM-1 engine architecture

as presented in Fig. 3. The DSP module is responsible for capturing an audio signal from a device or from a file and processing it. This stage may include normalization and frequency filtering depending on the input signal characteristics. Next, the signal is divided into observations with 25 ms window and 10 ms stepping, and for each observation a set of parameters is calculated in the form of a feature vector.

Feature extraction of the example speech signal represented as a spectrogram in Fig. 1 is performed not on the basis of the acoustic parameters of the individual phonemes represented by the fundamental frequency F_0, the frequencies of the individual formants, their bandwidths or amplitudes but by using the cepstrum of the signal represented in the Mel scale (mel-cepstrum). Cepstrum C_p, called spectrum of a spectrum, is a result of the inverse Fourier transform of the logarithm of the estimated signal spectrum:

$$C_p = \left| \mathcal{F} \left\{ \log \left(\left| \mathcal{F} \left\{ f(t) \right\} \right|^2 \right) \right\} \right|^2 . \tag{1}$$

The Mel scale is a perceptually motivated scale based on human hearing. It is approximately linear up to 1 kHz and logarithmic thereafter. In this way it represents the way we perceive frequencies. The formula for converting frequency f in Hz into a corresponding value in Mel scale is as follows:

$$\mathrm{Mel}\,(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right) . \tag{2}$$

For each observation LDA (Linear Discriminant Analysis) transformation is computed over Mel Frequency Cepstral Coefficients (MFCC) and filterbank parameters. MFCCs c_i are calculated from the log filterbank amplitudes m_j using the Discrete Cosine Transform:

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos \left[\frac{\pi i}{N} (j - 0.5) \right] , \tag{3}$$

where N is the number of filterbank channels. The DSP module also detects the speech signal by using a component called Voice Activity Detection (VAD). Only observations that contain the speech signal are passed onto the next module.

The decoder processes the feature vectors that characterize the individual observations generated by the DSP module in order to recognize the words that make up the utterance. The decoder is built upon a modified beam-search algorithm and operates over a recognition network containing words with imposed unigram probabilities. The decoding process involves finding the best match between a sequence of observations and a path in the recognition network. A set of hypotheses is selected, each of which matches the input observation string with certain probability.

The hypotheses generated by the decoder are subjected to a rescoring process performed by the Rescorer module. Its purpose is to evaluate individual hypotheses in terms of combined acoustic and linguistic probability and to select the most likely hypothesis. The ARM-1 Rescorer uses a 3-gram language model.

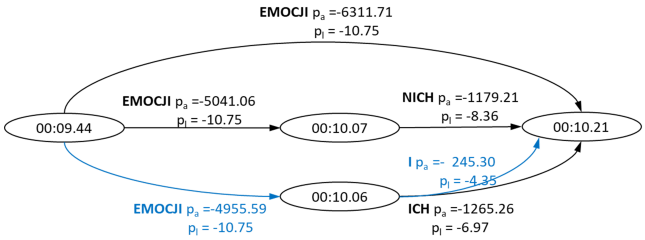


Fig. 4. Hypotheses lattice

Fig. 4 presents an example of a lattice with several hypotheses which is processed by the Rescorer. Each node represents a moment in time and each edge represents a word with a given acoustic and linguistict probability, where p_a and p_l are their logarithms. The best path selected by the Rescorer is marked in blue.

III. 3. 2. ARM-1 Adaptation for Media Monitoring

Application of the speech recognition technology to media monitoring, i.e. transcription of radio and television broadcast content, requires adaptation of the ASR methods to the characteristics of such content, its specific quality and technical features. Spontaneous speech, background noise, dialogues, sudden changes of speakers and overlap-

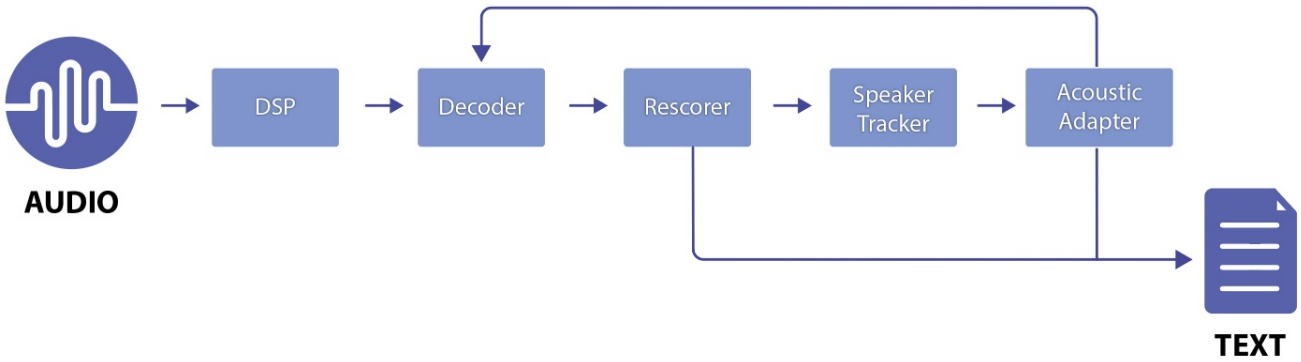


Fig. 5. ARM-1 engine with adaptation modules

ping sounds affect speech recognition results. There is also interference typical to radio and television broadcast content such as quiet background music played during new services or sounds (usually musical) marking the beginning of a program or an advertising block – the so-called jingles. On top of that there is a need to recognize foreign words that occur in an utterance in Polish and whose pronunciation may be affected by the context.

Adaptation of the ARM-1 engine to the above audio content characteristics included adaptation of acoustic and language resources and modification of the engine software.

The base acoustic model was built on the basis of a large corpus of recordings for Polish appropriately selected in terms of linguistic conditions, phonetic and acoustic properties of the language, the core of which was the JURISDIC database [26]. This set was gradually supplemented with additional recordings of dictated speech, parliamentary speeches, telephone conversations, court hearings, conferences and workshops. The variety of recordings made it possible to include speech of different nature and degree of spontaneity, as well as different acoustic characteristics in the model.

Prior to generating an acoustic model, all recordings had to be annotated taking into account all necessary linguistic phenomena and sound information to create a description including orthographic transcription, segmentation, marking of unintelligible fragments, fragments containing interferences, non-linguistic (filled pauses, repetitions, false starts) and physiological (laughter, breathing, grunting) events. Training set preparation is the most time-consuming part of the acoustic resources processing procedure.

For the adaptation of the acoustic model to the media monitoring requirements, an additional acoustic database was built containing recordings from several different television and radio channels with a total duration of about 55 hours. This set was also subject to the augmentation procedure. The adaptation process included: 1) "narrowing" the acoustic model to a set of words extracted from a corpus of press texts on the basis of frequency and expert knowledge, 2) modification of parameters (mean and variance of the probability distribution of individual components of the feature vector). The total duration of all acquired recordings was over 2100 h.

The baseline language model was generated with 20 GB of text consisting mainly of press texts from several newspapers and magazines published over several years, Wikipedia articles and transcriptions of parliamentary speeches. Phonetic transcription of radio and television recordings was used to adapt the language model. Furthermore, the model was adapted to include the set of words that was extracted from the acoustic model. Due to the fact that the language of the media changes very dynamically, it was necessary to implement the functionality enabling management of the ARM-1 engine dictionary on an ongoing basis, in particular to supplement the language resources with new words.

In order to increase the recognition performance, ARM-1 engine's functionality was extended with an unsupervised acoustic adaptation mechanism for the speaker's voice, the microphone used, the acoustic environment and the transmission method. This course of action was dictated by the nature of media content exhibiting high variability with respect to speaker and acoustic characteristics.

The adaptation required introduction of additional modules: Speaker Tracker and Acoustic Adapter as presented in Fig. 5. Speaker Tracker is responsible for detecting speaker changes and collecting statistical data for each speaker identified in the input signal. It operates on the parametrized audio samples and the hypothesis produced by the decoding phase. Speaker change detection relies on the comparison between GMM components and audio parameters corresponding to HMM states. The Speaker Tracker module accumulates observations associated with each visited HMM state in the best hypothesis. The observation accumulators are separate for each detected speaker. Such an approach enables parallel and independent acoustic adaptation of multiple speakers, e.g. during a dialogue.

The Acoustic Adapter module adapts the acoustic model to the speaker and estimates possible improvement resulting from the model adaptation. It operates on statistical audio parameters data collected for a given speaker and calcu-

lates a set of transformations for GMM components using Maximum Likelihood Linear Regression (MLLR). This step reduces the mismatch between observation features and the GMM components. Because the amount of adaptation data is very limited in this case, a regression class tree which groups GMM components is used [24].

After acoustic model adaptation the input data is decoded again provided that the estimated improvement exceeds a certain threshold. Since a large percentage of radio and TV recordings contain dialogues and are characterized by frequent speaker changes, in order to better exploit the potential of unsupervised adaptation, an utterance that was originally processed with a model of the speaker different from the one identified in the recording, is decoded again. The secondary decoding is performed in the same way as the primary one by the Decoder module but this time with the adapted models. The aforementioned mechanism is based on a simple predictor of the "profitability" of re-decoding, based on the change in error (variance) of the adaptive versus speaker-independent model for a given utterance. In this way recognition accuracy can be improved considerably if the utterance segmentation based on speaker change detection was performed correctly. In the case of rapid changes of the acoustic conditions and the speakers it is beneficial to decode the entire selected parts of the utterance again.

III. 4. DNN-based Approach

The next milestone in the development of automatic speech recognition was the use of neural network algorithms. The main idea behind artificial neural network application was to mimic the behavior of a human brain with layers of neurons.

Neural networks have quite a long history but their potential could be exploited only after the introduction of the so-called backpropagation method [27, 28]. This method enabled efficient and effective training of neural networks consisting of multiple layers of neurons and consequently enabled further development of neural networks. Several years later the so-called Deep Learning was introduced and is currently used virtually in every state-of-the-art solution based on machine learning algorithms. Deep neural networks based on the most modern network architectures are trained with huge amounts of data in order to have the ability to extract a lot of important information necessary to carry out recognition or classification tasks. Such high quality recognition models have a significant advantage over standard statistical methods. It may be difficult to analyze rationale behind specific decisions reached by a DNN-based solution for a classification problem but it is generally the case that each subsequent layer of neurons represents a higher level of abstraction in modeling a given phenomenon.

In the case of speech recognition two main paths can be distinguished for an application of deep neural networks: 1) replacing only the GMM model with a neural network, and 2) implementing the entire acoustic model with the neural network. In the first approach the HMM model is still used but it is combined with the neural network to form a hybrid acoustic model called DNN-HMM. This approach is still relatively common. However, it is worth noting that the introduction of neural networks in ASR technologies did not suddenly result in ideal recognition models that were immune to noise and could recognize speech with very high accuracy. Despite achieving much better results compared to the statistical models, deep neural network algorithms are constantly being improved and optimized to increase recognition quality. DNN-based ASR systems (including hybrid DNN-HMM systems) have been achieving high recognition accuracy for words that are pronounced clearly (e.g., dictated speech) without the presence of significant ambient noise. However, to date, recognition quality for incomplete, noisy or specific data leaves much to be desired.

The amount and quality of data used in the training process has a significant impact on the final quality of models. Without an adequate set of data it is impossible to create good quality recognition models based on DNN algorithms. Neural networks, especially the deep ones, are machine learning methods very sensitive to the amount of learning data.

In the task of speech recognition the most common types of neural networks used for many years have been recurrent neural networks (RNNs) [29–32] which enable models with high recognition accuracy. On the other hand, the state-of-the-art neural network architecture that is being successfully deployed in many different machine learning domains (including speech recognition) are Transformers neural networks [33, 34]. These networks have achieved an increased recognition accuracy but it should be pointed out that training them requires even more data than standard deep neural networks.

In addition to the development of the neural network architectures also the very approach to the ASR solution development has been verified. Currently, the most common approaches are hybrid and end-to-end solutions [35]. In the hybrid approach a speech recognition system consists of several independent main components which usually include an acoustic model, a pronunciation model and a language model. These components are created and trained separately. Additionally, in order to enable final recognition a decoding graph is built in which a search for the best transcription is performed. In contrast, the end-to-end approach is a system whose all parts are trained together to directly map the input sequence of acoustic features to a target output sequence in the form of, for example, words.

The ASR engine applied in media monitoring solutions which achieved the best results was based on a hybrid approach using DNN. The best result was reported for Estonian with WER $= 8\%$ for broadcast news. However, taking into account various tasks not limited to media monitoring, end-to-end solutions achieve results that are similar (and sometimes even better) to hybrid models. The results obtained

depend strictly on the train and test set used, e.g. for WSJ eval92 [36] the best result obtained so far is WER = 2.32% (TDNN+chain), while for the Librispeech [37] set the lowest WER is 1.4% (wav2vec 2.0 [38]). An overall comparison of several end-to-end and hybrid solutions for the Microsoft dataset and a proposal to combine these two approaches is presented in [39].

III. 5. DNN-based ARM-1 NG Engine

Development of the ARM-1 engine also followed the path described in the previous section. Despite the use of large learning sets, both acoustic and linguistic, the average recognition error for the GMM/HMM based ARM-1 engine exceeded 20%, much too high for the considered application to media monitoring.

The new ARM-1 NG (Next Generation) engine is based on a hybrid approach. It has been implemented with the use of Kaldi software tools [40, 41] for generating components based on the provided training data. Kaldi is an open source speech recognition toolkit intended for use by ASR researchers for building recognition systems. It supports modeling of context-dependent phones of arbitrary context lengths and all commonly used techniques that can be estimated using maximum likelihood.

The main component of ARM-1 NG, the acoustic model, is based on a hybrid DNN-HMM solution generated by Kaldi as a part of deep neural networks implementation. A number of implementations (nnet1, nnet2 and nnet3) provided by Kaldi are dedicated to creating custom acoustic models optimized for efficient training using distributed computing resources. For the implementation of the new ARM-1 NG acoustic model the nnet3 implementation was chosen or, to be more precise, its extended version “nnet3+chain”.

The model is based on the classical parameterization of the acoustic signal using MFCC and LDA. In addition, 3-fold stacking subsampling vectors [42] are used (Fig. 6), which reduces the number of observations three times (originally 100 samples/s in ARM-1). This approach is justified because nnet3 implementation already accounts for the temporal aspect.

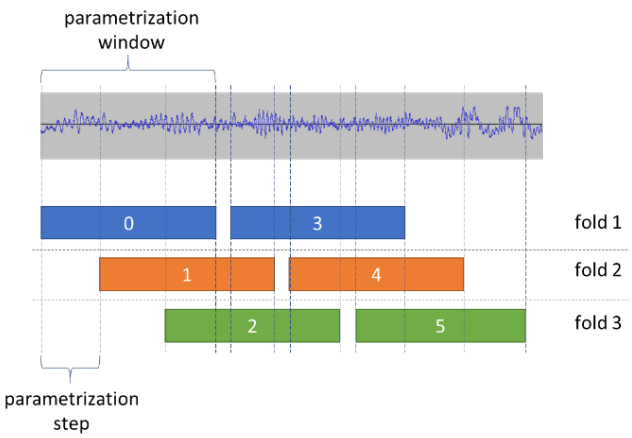


Fig. 6. 3-fold stacking subsampling operation

A phoneme-level recognition network is constructed using a phoneme N-Gram model computed from training data. The main difference with the ARM-1 engine is the change in the objective function used to search the recognition network. Originally, the cost was optimized at a single frame level, representing the log of acoustic similarity for each state of the HMM. In the chain model the log of phoneme sequence probability is optimized.

The change in observation frequency necessitated a change in the HMM network architecture. Originally, a minimum of 3 transitions were required to pass the HMM model. After the change the topology contains a state that must be reached at least once while all other states can be reached zero or more times. Fig. 7 illustrates this concept with the upper part showing an example of HMM network before change in the observation frequency and the lower part showing the network after the change.

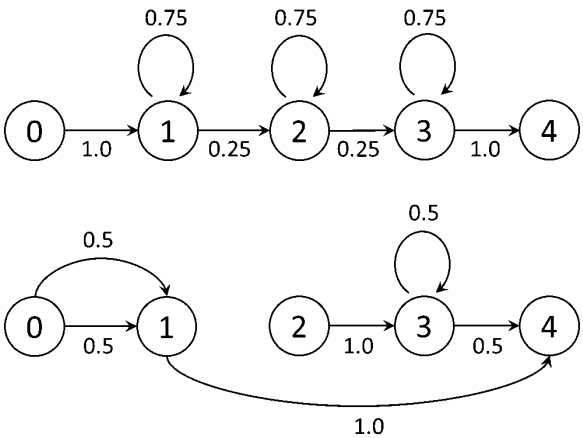


Fig. 7. HMM network architecture change

IV. RESULTS AND DISCUSSION

The ARM-1 engine performance was evaluated throughout its development using standard ASR evaluation metrics such as Correctness and Accuracy. In this Section we present results of tests performed using data sets constructed with broadcast media content in order to assess ARM-1’s effectiveness for media monitoring activities.

IV. 1. Evaluation Measures

Speech recognition results are evaluated by comparing an utterance, i.e., the manual transcription of the recording, with the transcription generated by the ARM-1 engine and determining the Levenshtein distance [43]. This metric is defined for two sequences of words as the minimum number of single word edit operations required to change one sequence

into the other. The edit operations include insertion, deletion and substitution.

Each word occurring in either sequence is classified as correctly recognized (Correct), incorrectly recognized (Substitution), deleted (Deletion) or inserted (Insertion). The number of words in each category is used to determine the Word Error Rate (WER):

$$\mathrm{WER} = \frac{S + D + I}{N} \,, \tag{4}$$

where S is the number of incorrectly recognized words, D is the number of words not recognized at all, I is the number of extra words that were recognized but do not have corresponding words in a given utterance and N is the number of words in the utterance. Speech recognition results can also be characterized with Correctness and Accuracy. The first metric shows what part of the utterance was correctly recognized:

$$\mathrm{Corr} = \frac{H}{N} \,, \tag{5}$$

where H is the number of correctly recognized words. Accuracy in addition to incorrectly recognized words takes into account also insertions:

$$\mathrm{Acc} = \frac{H - I}{N} = 1 - \mathrm{WER} \,. \tag{6}$$

Accuracy is more precise than Correctness since insertions can affect the quality of speech recognition results considerably especially for some ASR applications. In general, different types of recognition errors have different effect. For example, if recognition results are used to search for a given word phrase, Insertions may increase the number of false positives while Deletions may increase the number of false negatives.

Correctness of speech recognition results has a direct bearing on the effectiveness of media monitoring regardless of its purpose. However, the standard metrics are quite strict and do not take into account the character of the language. Polish is an inflectional language where one word can have as many as a dozen different grammatical forms. Recognizing a correct word but an incorrect form of this words counts as an error. Depending on the purpose of media monitoring activity, WER computed according to Eq. (4) could in fact be treated as an upper bound on the actual error rate.

IV. 2. Test Sets

Four sets of data were used to evaluate and compare speech recognition accuracy of the ARM-1 and ARM-1 NG engines. In the following description we refer to the two versions of the engine simply as ARM-1 and ARM-1 NG. The sets were created in the process of the system development for performance monitoring purposes. Each test set is described below.

1. **Set1** contains 13 radio news services broadcasted in 2015–16 by a number of local (PR Szczecin, PR Wrocław, Radio Plus Łódź) and national (Eska, ZET, Meloradio) radio stations. The total duration is about 1 hour 2 minutes. The shortest recording has 18 seconds, the longest 8 minutes.
2. **Set2** contains 11 radio and television programs (news services, interviews, journalistic programs) broadcasted in 2015–17 by the national radio (ZET, Tok FM) and TV stations (TVP, TVN, Polsat). The total duration is 2 hours 14 minutes. The length of the recordings varies from 3 to 27 minutes.
3. **Set3** contains 433 short news services broadcasted in 2015–16. The set was generated by manually cutting the "Set1" collection into segments based on the content type (speech, jingle, etc.) and speaker so that there is only one speaker in each segment. Furthermore, the segments were tagged based on speaker characteristics – professional speaker (announcer, journalist) or nonprofessional speaker (interview guest, passerby), and on acoustic conditions – studio or outdoor recording. The objective of dividing the recording set into fragments was to obtain segments that are uniform in terms of speaker and acoustic conditions to identify elements for which recognition results have high WER and to determine the cause. The majority of segments in terms of duration were recorded in the studio (about 75%) and were spoken by professional speakers (also about 75%). The longest segment has a duration of 70 s.
4. **Set4** contains 29 recordings of proceedings of the Polish Parliament dating back to 2019. The total duration is 48 minutes. The length of the recordings varies from about 20 seconds to 4 minutes. This set was assembled for PolEval 2019 [44] ASR task and as such can be used for comparing speech recognition results of various ASR systems.

IV. 3. Test Results

The tests were conducted for ARM-1 with and without the adaptation module, and for ARM-1 NG with the dictionary of two different sizes. The adaptation procedure has not been implemented in ARM-1 NG yet. ARM-1 was equipped with a dictionary of 470k words, while ARM-1 NG had two dictionaries with 100k and 500k words. It should be mentioned that all words spoken in Set1, Set2 and Set3 were present in all dictionaries. ARM-1 NG was equipped with a different dictionary than ARM-1 due to updates of language resources necessitated by changes in media vocabulary that take place over time.

Fig. 8 presents test results in the form of WER obtained for the first three test sets and for every version of the system. In general, ARM-1 NG achieved WER roughly twice as small as ARM-1 confirming the potential of applying DNN-

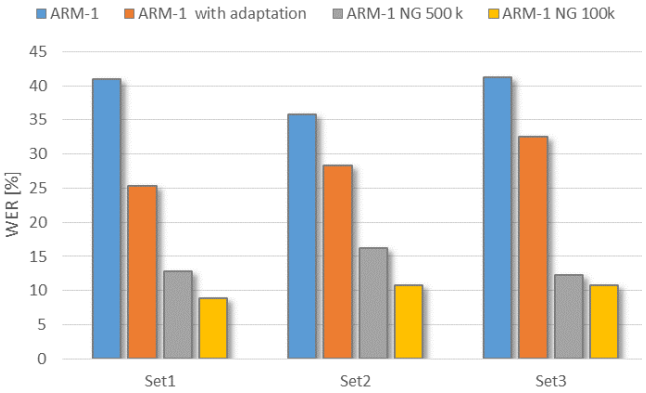


Fig. 8. Test results – WER comparison

based method. Speaker adaptation decreased WER by about 15% for Set1 and by several percent for Set2 and Set3.

The dictionary size affected ARM-1 NG performance in a slightly counterintuitive manner since lower WER was obtained with a smaller dictionary. However, it can be easily explained given the fact that all words spoken in the test sets were present in both dictionaries. Hence, using the smaller dictionary did not cause any out-of-vocabulary problems. On the other hand, the bigger dictionary offered more choices as to which word was spoken, increasing chances of selecting the wrong one.

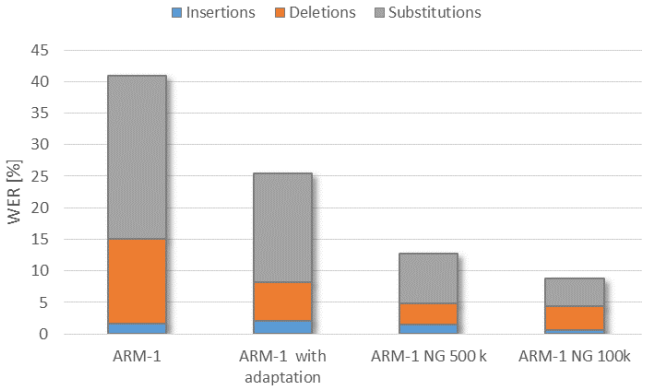


Fig. 9. Percentage of insertions, deletions and substitutions in the overall number of errors

Fig. 9 presents the percentage of each type of errors (Insertions, Deletions and Substitutions) in the overall error number made by each version of the ARM-1 engine for Set1. It can be observed that adaptation increased the number of Insertions slightly while using a smaller dictionary caused an increase in the number of Deletions and a decrease in the number of Insertions.

Set3 was used to compare recognition results obtained for studio and outdoor recordings, and for professional and nonprofessional speakers. As Fig. 10 shows, WER for studio recordings is much lower for all versions of ARM-1 engine than for outdoor recordings. One can observe that the

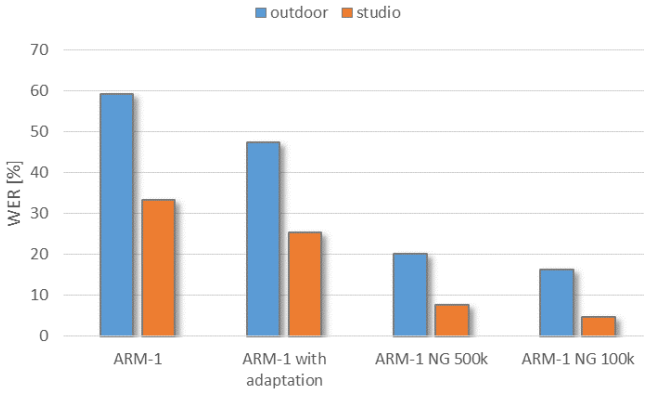


Fig. 10. WER comparison for various acoustic conditions

difference between the two results is much higher for the DNN-based system. The best WER for the studio category obtained with ARM-1 NG was 4.84% versus 16.32% for the outdoor category, while the best WER for the studio category obtained with ARM-1 was 25.24% versus 47.41% for the outdoor category.

Similarly, WER for the professional speaker group's recognition results is much lower than for the nonprofessional speaker group (Fig. 11). And also in this case the difference between the two types of speakers was larger for ARM-1 NG than for ARM-1. The best WER obtained with ARM-1 NG was 5.59% for professional speakers versus 14.6% for nonprofessional speakers. The best WER obtained with ARM-1 was 25.85% for professional speaker versus 46.32% for nonprofessional speakers. This behavior can be attributed to a smaller representation of outdoor and nonprofessional speakers recordings in the training set compared to studio and professional speakers category, respectively, and to higher sensitivity of the DNN-based approach to the amount and quality of training data.

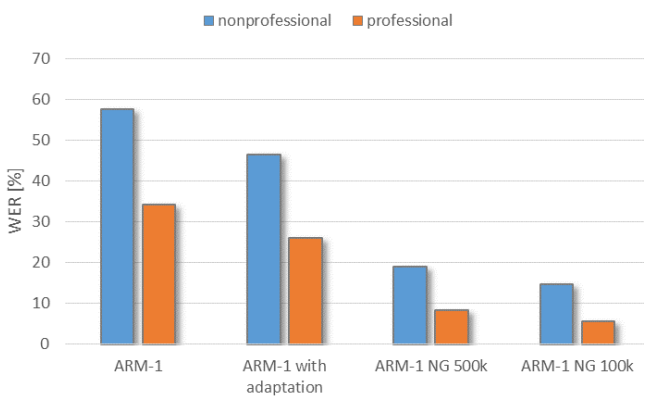


Fig. 11. WER comparison for two types of speakers

Set4 contained data that were different in character from the other test sets. First of all, recordings of the parliamentary speeches were made under uniform acoustic conditions with occasional background sounds caused by vocal reac-

tion of the audience. For most recordings there was only one speaker with the exception of short interruptions by the House Speaker leading the debate.

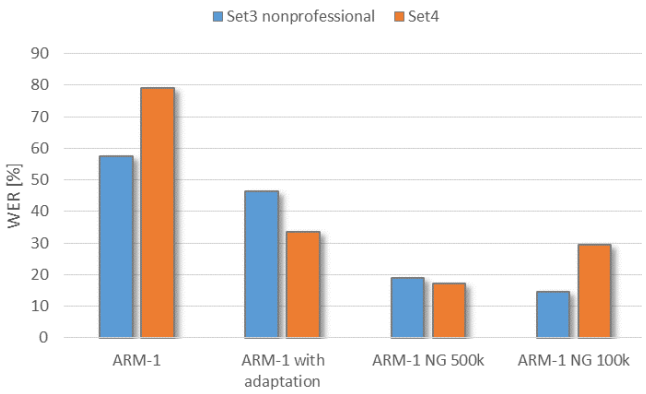


Fig. 12. WER comparison for parliamentary speeches

In general, WER was much higher for Set4 than for other test sets. There are several reasons for that. First of all, not all words spoken were present in the dictionaries used in the recognition process. Fig. 12 shows that WER obtained with ARM-1 NG was lower for the larger dictionary this time. Second of all, most speakers were nonprofessional speakers and their speeches was characterized by a high degree of spontaneity. We have compared results obtained for Set4 with the nonprofessional speaker category of Set3 as the test set closest in nature to Set4. The results obtained for both sets are generally comparable for ARM-1 NG with the 500k dictionary. The use of a smaller dictionary yielded higher WER due to an out-of-vocabulary problem. Speaker adaptation was able to lower WER quite considerably and worked better for Set4 most likely due to lower variability of the acoustic conditions.

The test results allowed us to draw a conclusion that the DNN-based system is much better suited for media monitoring, as expected. The WER value that does not exceed a few percent confirms that ARM-1 NG can be successfully applied in media monitoring regardless of its purpose.

It should be emphasized that it is important to keep updating the system's dictionary to reflect changes in the media language. This procedure involves not only adding new words but also removing words that become obsolete in order to control the dictionary size and decrease the chance of an incorrect recognition involving such an obsolete word. We are yet to determine the potential of applying the speaker adaptation procedure to the DNN-based system.

IV. 4. ASR Systems Comparison

ARM-1 engine was one of the competitors taking part in the PolEval 2019. Hence, we can present the comparison of its performance with other systems based on the PolEval ASR task results reported in [44]. The test set used in the competition is described as Set4 in the previous Section. However, the acoustic model used by ARM-1 for PolEval task was different from the model used for tests described in the previous Section. Consequently, test results reported for PolEval and are different from the results presented for Set4 in the previous Section. In the following description we first describe PolEval results and then comment on the results obtained by ARM-1 and ARM-1 NG for Set4 with different acoustic models.

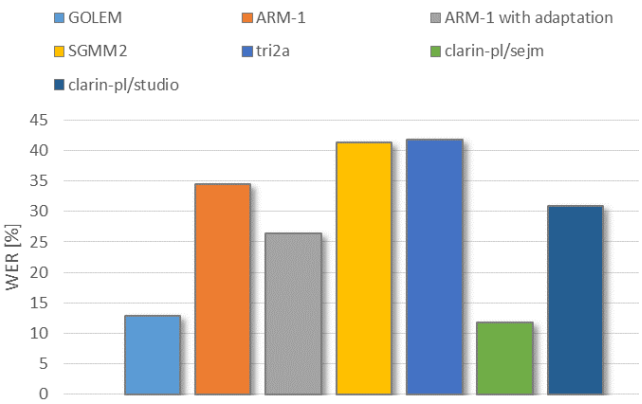


Fig. 13. PolEval ASR task result comparison

Fig. 13 presents a comparison of the results obtained by the ASR systems taking part in the competition, reported in [44]. Beside ARM-1 there were three more systems competing: GOLEM, SGMM2 and tri2a, all based on Kaldi project. All competing systems used GMM-based acoustic models. Results obtained by two more systems were added by the organizers for comparison: clarin-pl/sejm and calrin-pl/studio, both of which were neural network based. Among all systems, GOLEM and calrin-pl/sejm used models trained only with corpora selected by the organizers that included: 1) CLARIN-PL speech corpus, 2) PELCAR parliamentary corpus, 3) a collection of 97 hours of parliamentary speeches and 4) Polish Sejm Corpus for language modeling. All other systems used models that were trained on other data as well.

Among GMM-based systems the best results were obtained by GOLEM. The WER equal to 12.8 was quite close to WER obtained by the neural network based clarin-pl/sejm system which was 11.8. Both systems' acoustic models were trained on the fixed set of corpora selected by the organizers. We can only speculate that this set was dominated by recordings of parliamentary speeches and that the resulting acoustic models can be characterized as models dedicated to this type of recordings. The acoustic model used by ARM-1 engine for the competition was trained on a set in which recording of parliamentary speeches constituted only about 10% of all recordings in terms of total duration. Other training corpora included recordings of dictated speech, radio and TV programs and phone conversations. The results obtained by ARM-1 with the acoustic model trained mainly for media monitoring were much worse than results obtained with the

model used for PolEval, both with and without adaptation: 79.02 vs 34.49 and 33.69 vs 26.01, respectively.

ARM-1 NG engine's performance can be compared with that of clarin-pl/studio since they both use neural network based acoustic models trained on corpora not limited to the fixed set selected for PolEval. The lowest WER obtained by ARM-1 NG for PolEval test set was 17.11 which is lower than WER obtained by calrin-pl/studio equal to 30.9.

V. SUMMARY AND CONCLUSIONS

Our experience with GMM/HMM and DNN-based methods for speech recognition and their application in media monitoring allows us to conclude that the neural network based approach was needed to achieve accuracy necessary for this area to benefit from the automation of speech recognition. The DNN-based acoustic model trained on a large dedicated corpus yielded results that were sufficient for media monitoring, i.e. it achieved WER below 10%. However, there is still room for improvement in a DNN-based approach in general (e.g. extending DNN to the entire speech processing pipeline) and in adjusting the approach to media content characteristics in particular.

There are a number of future work directions we want to pursue. They include improvements at various stages of the speech signal recognition process from signal pre-processing including parametrization, through speaker diarisation, to speech recognition results post-processing including automatic correction and inserting punctuation.

We plan to test other methods of speech feature extraction (e.g. LPC, PLP, RASTA) [45] and an additional preprocessing module based on empirical signal decomposition. The empirical signal decomposition makes it possible in practice to extract significant signal components located in a specific frequency band by reducing noise, background sounds and unwanted sounds. We plan to conduct research on the use of Fourier Decomposition [46], Multivariate Variational Mode Decomposition [47], Enhanced Empirical Mode Decomposition [48], which are used in various fields for the decomposition of nonlinear, non-stationary and heavily noisy signals [48–52].

We plan to explore another area of ASR application, namely digital humanities, where automatic speech recognition enables audio content analysis. We expect this area to pose a similar challenge to media monitoring due to the diversity of content and the need to adjust the system to content characteristics.

In the long term, the use of more advanced language models, modeling of acoustic and prosodic phenomena of speech should further improve the performance of automatic speech recognition systems. However, the key to achieving the ability to recognize continuous and spontaneous speech will most likely be automatic understanding of the utterance meaning, i.e. semantic analysis. Also pragmatic analysis, i.e. understanding and interpreting utterances depending on the context, will be very helpful. The realization of these tasks, especially the application of semantic and pragmatic analysis, seems to be rather distant future. Meanwhile, specialized systems are being developed for application in particular areas of life.

Acknowledgment

Work presented in this paper was partially supported by the Polish Center for Research and Development in the following projects DOBR/0008/R/ID1/2013/03, DOB-BIO6/22/133/2014, POIR.01.01.01-00-1532/15-00 and POIR.04.01.04-00-0055/18.

References

[1] J. Jamroży, E. Kuśmierek, M. Lange, M. Owsianny, *Przetwarzanie dźwięku i obrazu materiałów radiowo-telewizyjnych – wyszukiwanie informacji multimedialnej*, [In:] *Postępy badań w inżynierii dźwięku i obrazu – Nowe trendy i zastosowania technologii multimedialnych*, Ed. B. Kostek, Akademicka Oficyna Wydawnicza EXIT, 194–225 (2019).

[2] J. Jamroży, M. Lange. M. Owsianny, M. Szymański, *ARM-1: Automatic Speech Recognition Engine*, [In:] *Proc. of the PolEval 2019 Workshop*, Eds. M. Ogrodniczuk, Ł. Kobyliński, 79–88 (2019).

[3] C. Mazurek, *Digital Humanities – Challenges for Humanities in the Digital Society Era – Foreword*, Computational Methods in Science and Technology **24**(1), 5–6 (2018).

[4] G. Rigoll, *The ALERT system: advanced broadcast speech recognition technology for selective dissemination of multimedia information*, IEEE Workshop on Automatic Speech Recognition and Understanding, 301–306 (2001).

[5] A. Znotins, K. Polis, R. Dargis, *Media monitoring system for Latvian radio and TV broadcasts*, Proc. of the 16^{th} Annual Conference of the International Speech Communication Association (INTERSPEECH), 732–733 (2015).

[6] M.S. Maučec, A. Žgank, *Speech recognition system of Slovenian broadcast news*, Speech Technologies, 221–236 (2011).

[7] H. Meinedo, *Audio Pre-Processing and Speech Recognition for Broadcast News*, PhD Thesis, IST, Technical University of Lisbon, Lisbon, Portugal (2008).

[8] H. Meinedo, J. Neto, *A stream-based audio segmentation, classification and clustering pre-processing system for broadcast news using ANN models*, 9^{th} European Conference on Speech Communication and Technology, 237–240 (2005).

[9] H. Meinedo, A. Abad, T. Pellegrini, J. Neto, I. Trancoso, *The L2F Broadcast News Speech Recognition System*, [In:] *Proc. of FALA*, 93–96 (2010).

[10] H. Meinedo, D. Caseiro, J. Neto, I. Trancoso, *Computational Processing of the Portuguese Language*, 6^{th} International Workshop PROPOR, Springer, 9–17 (2003).

[11] H. Meinedo, N. Souto, J. Neto, *Speech recognition of broadcast news for the European Portuguese language*, Automatic Speech Recognition and Understanding (2001).

[12] T. Alumäe, O. Tilk, *Automatic Speech Recognition System for Lithuanian Broadcast Audio*, Frontiers in Artificial Intelligence and Applications **289**: Human Language Technologies – The Baltic Perspective, 39–45 (2016).

[13] M. Sazhok, R. Selukh, D. Fedorin, O. Yukhimenko, V. Robeyko, *Automatic Speech Recognition for Ukrainian Broadcast Media Transcribing*, Control Systems and Computers 46–57 (2019).

[14] I. Demiros, H. Papageorgiou, V.Antonopoulos, Vassilios, A. Pipis, A. Skoulariki, *Media Monitoring by Means of Speech and Language Indexing for Political Analysis*, Journal of Information Technology & Politics **5**, 133–146 (2008).

[15] B. Gerazov, Z. Ivanovski, *Towards a System for Automatic Media Transcription in Macedonian*, 28^{th} Telecommunications Forum (TELFOR), 1–4 (2020).

[16] J. Neto, H. Meinedo, M. Viveiros, *A media monitoring solution*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1813–1816 (2011).

[17] R. Menon, A. Saeb, H. Cameron, W. Kibira, J. Quinn, T. Niesler, *Radio-browsing for developmental monitoring in Uganda*, [In:] Proc. of ICASSP, 5795–5799 (2017).

[18] M. Stadtschnitzer, *Robust Speech Recognition for German and Dialectal Broadcast Programmes*, PhD Thesis, University of Bonn, Bonn, Germany (2018).

[19] R. Safarik, J. Nouza, *Unified Approach to Development of ASR Systems for East Slavic Languages*, [In:] *Statistical Language and Speech Processing, Lecture Notes in Computer Science* **10583**, Eds. N. Camelin, Y. Estève, C. Martín-Vide, 193–203, Springer (2017).

[20] T. Alumäe, O. Tilk, A. Ullah, *Advanced Rich Transcription System for Estonian Speech*, Baltic HLT (2018).

[21] H. Meinedo, D. Caseiro, J. Neto, I. Trancoso, *AUDIMUS.MEDIA: A Broadcast News Speech Recognition System for the European Portuguese Language*, [In:] *Computational Processing of the Portuguese Language* **2721**, Eds. N.J. Mamede, I. Trancoso, J. Baptista, M. das Graças Volpe Nunes, Berlin, Springer, 9–17 (2003).

[22] Y. Miao, M. Gowayyed, F. Metze, *EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding*, [In:] *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 167–174 (2015).

[23] S. Grocholewski, *Statystyczne podstawy systemu ARM dla języka polskiego*, Wydawnictwo Politechniki Poznańskiej (2001).

[24] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, P. Woodland, *The HTK book*, Cambridge University Engineering Department **3** (2002).

[25] G. Demenko, M. Szymański, R. Cecko, M. Lange, K. Klessa, M. Owsianny, *Development of large vocabulary continuous speech recognition using phonetically structured speech corpus*, Proc. of the 17^{th} International Congress of Phonetic Sciences (ICPhS XVII), A86–A98 (2011).

[26] G. Demenko, S. Grocholewski, K. Klessa, J. Ogórkiewicz, A. Wagner, M. Lange, D. Śledzinski, N. Cylwik, *LVCSR Speech Database – JURISDIC*, New Trends in Audio and Video/Signal Processing Algorithms, Architectures, Arrangements, and Applications SPA, 67–72 (2008).

[27] S. Linnainmaa, *The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors*, Master's Thesis (in Finnish), Univ. Helsinki (1970).

[28] D.E. Rumelhart, G.E. Hinton, R.J. Williams, *Learning representations by back-propagating errors*, Nature **323**, 533–536 (1986).

[29] A. Graves, A. Mohamed, G. Hinton, *Speech recognition with deep recurrent neural networks*, 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 6645–6649 (2013).

[30] H. Sak, A. Senior, F. Beaufays, *Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition* (2014).

[31] A. Amberkar, P. Awasarmol, G. Deshmukh, P. Dave, *Speech Recognition using Recurrent Neural Networks*, 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), 1–4 (2018).

[32] T.S. Vandhana, S. Srivibhushanaa, K. Sidharth, C.S. Sanoj, *Automatic Speech Recognition using Recurrent Neural Network*, International Journal of Engineering Research & Technology (IJERT) **9**(8) (2020).

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, *Attention Is All You Need*, Advances in Neural Information Processing Systems (NIPS) **30** (2017).

[34] L. Dong, S. Xu, B. Xu, *Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition*, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5884–5888 (2018).

[35] S. Wang, L. Guanyu, *Overview of end-to-end speech recognition*, Journal of Physics: Conference Series **1187**(5), IOP Publishing (2019).

[36] D.B. Paul, J.M. Baker, *The design for the Wall Street Journal based CSR corpus*, [In:] *Proc. of the workshop on Speech and Natural Language. Association for Computational Linguistics*, 357–362 (1992).

[37] V. Panayotov, G. Chen, D.Povey, S. Khudanpur, *Librispeech: An ASR corpus based on public domain audio books*, [In:] *Proc. Int. Conf. Acoust., Speech, Signal Process*, 5206–5210 (2015).

[38] Y. Zhang, J. Qin, D. Park, W. Han, C. Chiu, R. Pang, Q. Le, Y. Wu, *Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition* (2020).

[39] J.H. Wong, Y. Gaur, R. Zhao, L. Lu, E. Sun, J. Li, Y. Gong, *Combination of End-to-End and Hybrid Models for Speech Recognition*, Proc. of InterSpeech, 1783–1787 (2020).

[40] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, K. Vesel, *The Kaldi speech recognition toolkit*, IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (2011).

[41] D. Povey, X. Zhang, S. Khudanpur, *Parallel training of DNNs with Natural Gradient and Parameter Averaging*, arXiv: 1410.7455v4 (2015).

[42] H. Sak, A. Senior, K. Rao, F. Beaufays, *Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition*, Proc. of InterSpeech 2015, 1468–1472 (2015).

[43] V. Levenshtein, *Binary codes capable of correcting deletions, insertions, and reversals*, Soviet Physics Doklady **10**, 707–710 (1965).

[44] D. Koržinek, *Results of the PolEval 2019 Shared Task 5: Automatic Speech Recognition Task*, [In:] *Proc. of the PolEval 2019 Workshop*, Eds. M. Ogrodniczuk, Ł. Kobyliński, 73–78 (2019).

[45] K.R. Ghule, R.R. Deshmukh, *Feature extraction techniques for speech recognition: A review*, International Journal of Scientific & Engineering Research **6**(5), 143–147 (2015).

[46] P. Singh, S.D. Joshi, R.K. Patney, K. Saha, *The Fourier decomposition method for nonlinear and non-stationary time series analysis*, Proc. of the Royal Society A: Math., Phys. and Eng. Sci. **473**, 20160871 (2017).

[47] N. Rehman, H. Aftab, *Multivariate Variational Mode Decomposition*, IEEE Transactions on Signal Processing **67**(23), 6039–6052 (2019).

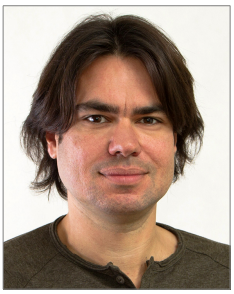
[48] Y. Hu, F. Li, H. Li, C. Liu, *An enhanced empirical wavelet transform for noisy and non-stationary signal processing*, Digital Signal Processing **60**, 220–229 (2017).
[49] P. Cao, H. Wang, K. Zhou, *Multichannel Signal Denoising Using Multivariate Variational Mode Decomposition With Subspace Projection*, IEEE Access **8**, 74039–74047 (2020).
[50] P. Kuwalek, B. Burlaga, W. Jesko, P. Konieczka, *Research on methods for detecting respiratory rate from photoplethysmographic signal*, Biomedical Signal Processing and Control **66**, 102483 (2021).
[51] P. Kuwalek, *Estimation of Parameters Associated With Individual Sources of Voltage Fluctuations*, IEEE Transactions on Power Delivery **36**(1), 351–361 (2021).
[52] P. Singh, A. Singhal, B. Fatimah, A. Gupta, S.D. Joshi, *AFMNS: A Novel AM-FM Based Measure of Non-Stationarity*, IEEE Commun. Lett. **25**(3), 990–994 (2021).

Robert Cecko is head of the New User Interface Technologies Department at Poznan Supercomputing and Networking Center. He obtained his MSc in Computer Science from Poznan University of Technology. Since then he worked as a computer system analyst at the Institute of Computer Science at Poznan University of Technology and then at Poznan Supercomputing and Networking Center as a computer systems analysis specialist. He was a leader of a development team for a number of projects related to automatic speech recognition and its applications in various areas including media monitoring.

Jerzy Jamroży obtained his MSc degree in Computer Science from Poznan University of Technology in 1998. Since 2003, he has been working at Poznan Supercomputing and Networking Center as a programmer and computer systems analyst. Since 2009, he has been involved in the implementation of the ARM-1 engine as well as tools and services based on the engine.

Waldemar Jęśko received the MSc degree in Electrical Engineering from Poznan University of Technology in 2018. He is currently a PhD student at the Faculty of Computing and Telecommunications at Poznan University of Technology. He works at Poznan Supercomputing and Networking Center in the New User Interface Technologies Department. His current research interests include neural networks, speech recognition and speech signal processing, in particular in the field of recognition of disrupted speech.

Ewa Kuśmierek obtained her PhD in Computer Information Sciences from the University of Minnesota in Minneapolis, USA in 2004. She currently works as a computer system analyst at Poznan Supercomputing and Networking Center in the New User Interface Technologies Department. She has participated in a number of research and development projects related to speech recognition and development of ARM-1 engine and its applications in various areas including media monitoring. She is a member of DARIAH-PL Music Information Retrieval Working Group whose objective is to develop a digital platform leveraging music information retrieval and analysis tools to facilitate interdisciplinary studies in musicology.



Marek Lange works as a computer system analyst at Poznan Supercomputing and Networking Center specializing in automatic speech recognition research and development. Prior to joining PSNC, he worked as a software engineer at PPN (Adam Mickiewicz University Foundation). Marek Lange received his MSc degree from Poznan University of Technology in Poland. He lives with his cat in Poznan, Poland. An incurable black coffee advocate and lifelong never-finished project creator.

Mariusz Owsianny is a physicist and acoustician, experienced in research work and computing, author of 25 research publications in physics and acoustics, as well as phonetics and speech technology. He worked as a sound engineer in the Polish Radio, acoustic specialist in the Phonetic Acoustics Department of the Institute of Fundamental Technological Research PAS and for international hearing aid companies. Currently he is a member of Laboratory of Integrated Speech and Language Processing Systems and takes part in activities related to speech analysis, speech processing, speech and speakers' recognition. He has participated in the development of a new version of the "Polish Numerical and Verbal Auditory Tests and Auditory Training Tests". Since 1994 he has been Secretary of the Polish Phonetic Association.

 DOI:10.12921/cmst.2021.0000015