DOI https://doi.org/10.15407/usim.2019.06.046
UDC 004.934

M.M. SAZHOK, Ph.D. (Eng.), head of the department, International Research and Training Center for Information Technologies and Systems NAS and MES of Ukraine, Glushkov ave, 40, Kyiv, 03187, Ukraine,
sazhok@gmail.com

R.A. SELIUKH, Research Associate, International Research and Training Center for Information Technologies and Systems NAS and MES of Ukraine, Glushkov ave, 40, Kyiv, 03187, Ukraine,
vxm112@gmail.com

D.YA. FEDORYN, Research Associate, International Research and Training Center for Information Technologies and Systems NAS and MES of Ukraine, Glushkov ave, 40, Kyiv, 03187, Ukraine,
dmytro.fedoryn@gmail.com

V.V. ROBEIKO, Research fellow, International Research and Training Center for Information Technologies and Systems NAS and MES of Ukraine, Glushkov ave, 40, Kyiv, 03187, Ukraine,
valya.robeiko@gmail.com

O.A. YUKHYMENKO, Research fellow, International Research and Training Center for Information Technologies and Systems NAS and MES of Ukraine, Glushkov ave, 40, Kyiv, 03187, Ukraine,
enomaj@gmail.com

AUTOMATIC SPEECH RECOGNITION FOR UKRAINIAN BROADCAST MEDIA TRANSCRIBING

A set of speech recognition techniques that allow for Ukrainian broadcast monitoring are covered: speech-to-text conversion; speaker diarization and recognition; text perception enhancement; multilingual aspects. The experimental results are presented and discussed.

Keywords: *speech, speech signal, analysis, recognition, understanding, synthesis.*

Introduction

Broadcast media is mainly a source of audio and, if applicable, video information. Ways the audience consumes broadcast media in the digital age have been transformed and transcription is the answer for the audience's many needs. The transcribed broadcast is searchable, transcripts serve the millions of people around the globe who are either deaf or hard of hearing, transcripts support social media and help creating new content, moreover, growing media monitoring companies aim to analyze as more as possible broadcast data and transcribing automation is crucial for them.

When applying automatic speech recognition for acquiring broadcast media transcripts we face the following issues:

- input speech signal is recorded in different including rather adverse acoustical conditions;
- speaker voice individuality modeling necessity;
- presence of more than one language in speech;
- capability for further automatic processing of the speech recognition result.

Hence, the target speech-to-text conversion system should be invariant to the wide range of noises and distortions introduced by equipment and compression algorithms. The system should adapt to speaker acoustical peculiarities and spontaneous changing of language. In turn, the succeeding use of speech recognition results, reveals new tasks related to presenting the recognized text in a convenient way for both human perception and subsequent automatic processing. This refers to availability of punctuation and marks of transition of speech from one person to another.

State-of-art

Nowadays, the dominating approach in speech-to-text technology is HMM-GMM-DNN, which is a combination of three basic methods. First of them, Hidden Markov Models (HMM), is a generative model that directly implements the principle of "analysis by synthesis" in order to represent non-linear temporal deformations of speech. The second approach is Gaussian Mixture Model (GMM) that allows for approximating of areas in the acoustic feature space where phonemes are observed. Finally, by means of Deep Neural Network (DNN) the quality of acoustical approximation is refined.

Representation of the relationship between HMM states and the acoustic input by increasing the number of diagonal Gaussians in GMM is effecient for speech signal self-segmentation at initial training stages. Despite advantages, GMMs are statistically inefficient for modeling data that lie on or near a non-linear manifold in the data space, which is typical for speech signal classes [1–3]. For example, modeling the set of points that lie very close to the surface of a sphere requires a few parameters using an appropriate model class, however, a very large number of Gaussians is required. Furthermore, a large number of Gaussians with independently parametrized means may result to local generalizations and drastic parameter increasing is required when changes are even caused by a small number of factors. In contrast to GMM, each parameter of a deep learning model is constrained by a large fraction of the data so is less amendable to local generalization. Moreover, DNNs benefit from exploiting multiple frames of input coefficients since no decorrelated input is required and, finally, DNN training is easier to parallelize.

Speech signal temporal non-linearity is modeled by generating of hypothetically valid model signals, their comparing with input signal and discarding wasteful model signal hypotheses. The HMM approach is an efficient way to generate model speech signals in accordance to the speech pattern hierarchy. For each variant of the word pronunciation, its acoustical model is composed from basic HMMs representing parts of context-dependent phoneme (senone states). Compositions of word models generate model word sequences where each word is considered in a certain lexical context. Thus word transitions are softly restricted in accordance to *n*-gram model, which parameter are being estimated by the domain-related text corpus.

The means to specify and operate generative models use representation of Weighted Finite State Transducers (WFST). Within the framework of this mathematical apparatus the formulated techniques to combine, compose and optimize generative models results in comprehensive and versatile construction for speech-to-text systems. An open source OpenFST toolkit allows for constructing generative models in terms of transducers without being distracted by the details of implementing optimization algorithms for the models [4, 5].

The application of the above mentioned approaches with the use of appropriate tools has already made it possible in Ukraine to create a series of basic speech information technologies that have been tested, in particular, for the basic search of segments of TV and radio broadcast by text [6, 7]. In this paper we represent the development of systems that move towards an efficient transcribing of the Ukrainian broadcast media to meet many needs of individuals and companies that consume and analyze the extracted content. Firstly, we describe the basic speech-to-text system architecture, then we focus on data preparation, parameter estimation, recognition result post-processing and multilingual issues illustrated with experimental results and broadcast media transcription system presentation.

General structure

Fig. 1 illustrates the architecture of speech-to-text conversion for the broadcast media transcribing. A recognition component, the actual recognizer, receives a speech signal extracted form media information at the input and, at the output, referring to a data and knowledge base (D&KB) produces a recognition response.

Recognizer receives Input Speech Signal extracted from a media file. When passing through the Speech Activity Detector (SAD), the signal is segmented by speech presence or absence [8]. For each segment where speech is detected, Preprocessor converts the signal into the feature space based on mel-frequency cepstral coefficients with mean value subtraction supplemented with the *i*-vecror in accordance to the speaker adaptation technique (SAT) [9]. The latter allows also for completing the speaker diarization procedure that will be described below [10]. Decoder estimates the similarity criteria value for all valid model signal hypotheses given the input signal, which is memorized in the Dynamic Programming graph referred as lattice. To speed-up the decoding process, on decoding stage, the lexical context is limited to bigrams and most frequent trigrams included to Language Model (LM). To account the influence of broader lexical context, the lattice might be rescored, i.e. a language model based on *n*-gram, $n>3$, is re-applied for the decoded lattice. In Postprocessor the result of decoding (or rescoring) is analysed and transformed to the final textual form that is Recognition Response. Thus, Recognizer output is one or, in case of multi-decision, more word sequence hypotheses supplemented with estimations of beginning, length, confidence and speaker identity for each word. Optional outcomes of Postprocessor are restored punctuation, abbreviation and digital number representation [11, 12], which is aimed to facilitate the human perception of the extracted text and prepare the recognition result for further processing by automatic means.

The described Recognizer structure mentions that speech signal is processed as it arrives until SAT application. In turn, decoding proceeds starting right after the *i*-vector is extracted for the entire segment with speech detected.

D&KB parameters are estimated on speech and text corpora by means of training modules [13]. These modules allow for estimating parameters for models of speech patterns related to different Recognizer's units.

Research and development

For each pattern type modeling, the domain-specific speech or text corpus is used. Among particular criteria the corpus must follow are: natural languages, topics, expected distribution for speakers and speech styles, acoustical surround and information transmission channel properties.

So, about 500 hours of annotated Ukrainian broadcast records were used to estimate the parameters for acoustic phoneme models, speech activity detection and speaker adaptation.

Fig. 2 illustrates how an expert can correct a broadcast episode segment using the Transcriber

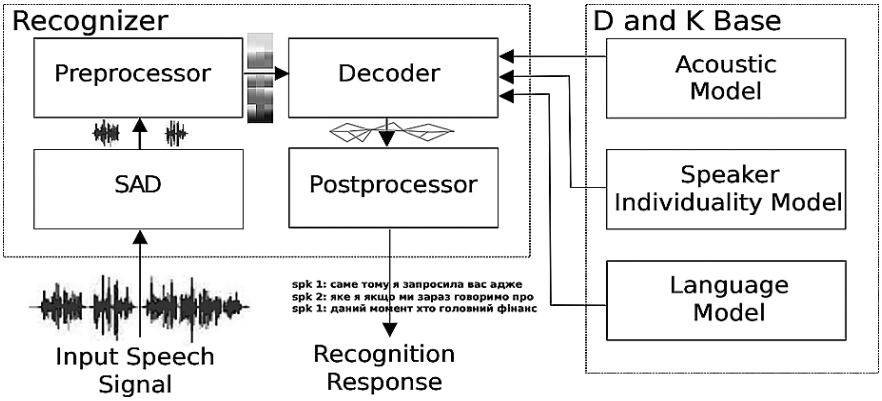

Fig. 1. Diagram of speech-to-text conversion architecture

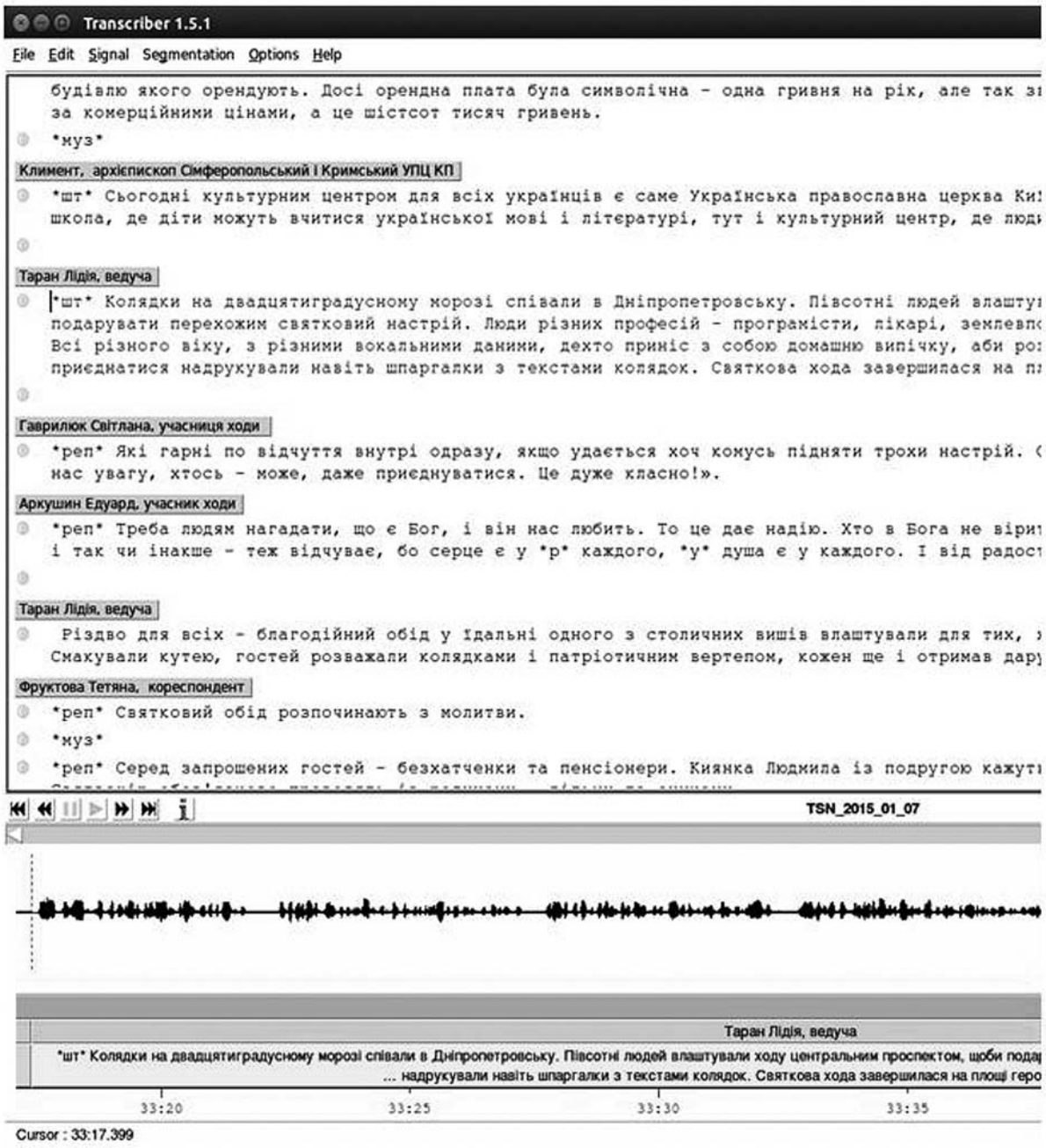

Fig. 2. Broadcast episode annotation

open source software. The expert inserts some derived from Ukrainian labels to mark speaker identity, language change and segments with lower SNR, including background noise, *шт*, and off-studio records, *реп*, segments with non-speech noise like music, *муз*, and publicity segments are labeled as well. On lexical level, abruptions and flufs are annotated. Therefore, this speech corpus is applicable for wider research range then we present in this paper.

To build a language model for word sequence restriction we collected a text corpus using over 2GB

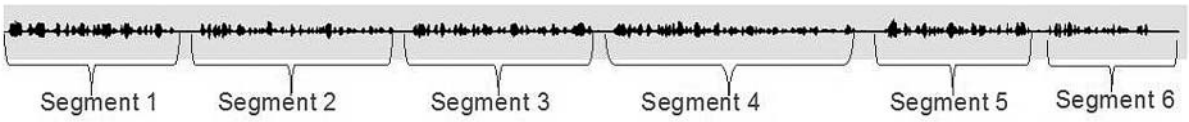

Fig. 3. Segments with detected speech activity

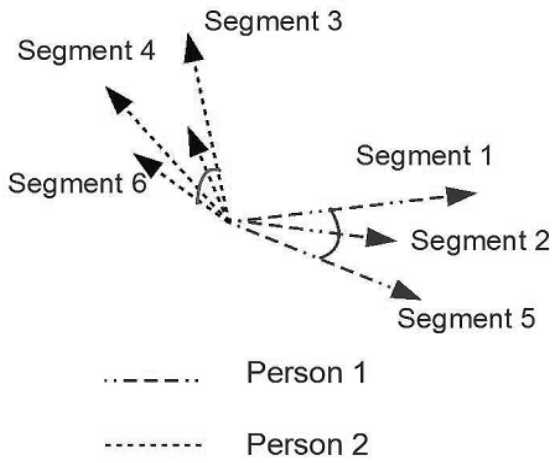

Fig. 4. Speaker diarization

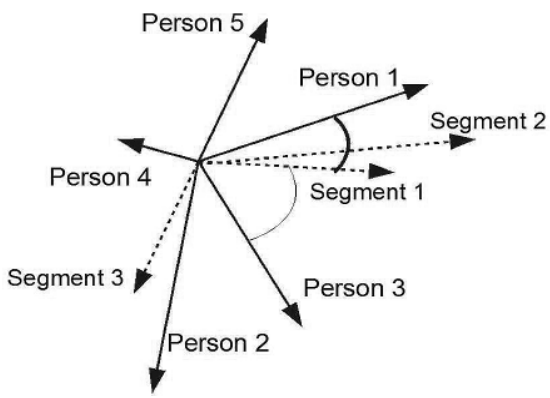

Fig. 5. Speaker identification

of text obtained from the Internet, particularly, using a subset of the Ubertext corpus.

To prepare recognition pronunciation dictionaries we use available supervised dictionaries [14, 15]. To estimate parameters for the automatic grapheme-to-phoneme we used the Sequitur converter based on Weig-hted Finite State Automaton. This allowed for predicting one or more pronunciations for words contained in the text corpus and not seen in the supervised dictionaries. However, this approach is not applicable for generating pronunciations by numbers in digit rather than spelled form. Therefore, all numbers are spelled in the text corpus by means of the rule-based symbol sequence-to-sequence converter [12].

Further processing of the speech-to-text result is carried out both by the human and by automatic means. One of the further options for processing the speech recognition result is to manually correct the errors in the received text. A human is interested in reducing the editions that must be made to obtain the final transcript. The sources of errors include incorrectly recognized as well as missing or inserted words, possible misspellings in the recognition vocabulary, editing of punctuation marks and speaker identity. In turn, one of the important goals of automatic word processing is the need to work with numbers.

Before restoring punctuation marks and converting word sequences to numbers, a speaker turn changing detection, i.e., speaker diarization, is accomplished. We assume that when the turn to speak is transferred to another party, the next thought is expressed, so a new sentence begins. Cases when other conversation parties can continue or end the sentence are not considered.

After the speaker diarization, we recover the numbers from the word sequences, and then restore the punctuation. This is due to the fact that the parameters of the punctuation model are more appropriate to evaluate on texts containing numeric expressions rather than word sequences. Otherwise, before evaluating the parameters, it would be necessary, in the training text sample, to convert the numeric expressions to a word sequence, which is not a completely solved problem, in particular for highly inflective languages and Ukrainian is among them. The speaker diarization goal is to answer the question who and when is speaking by the speech signal. For broadcast transcribing, the number of participants in the conversation is estimated automatically since it is unknown in advance. Afterwards, the obtained speech signal segmentation by

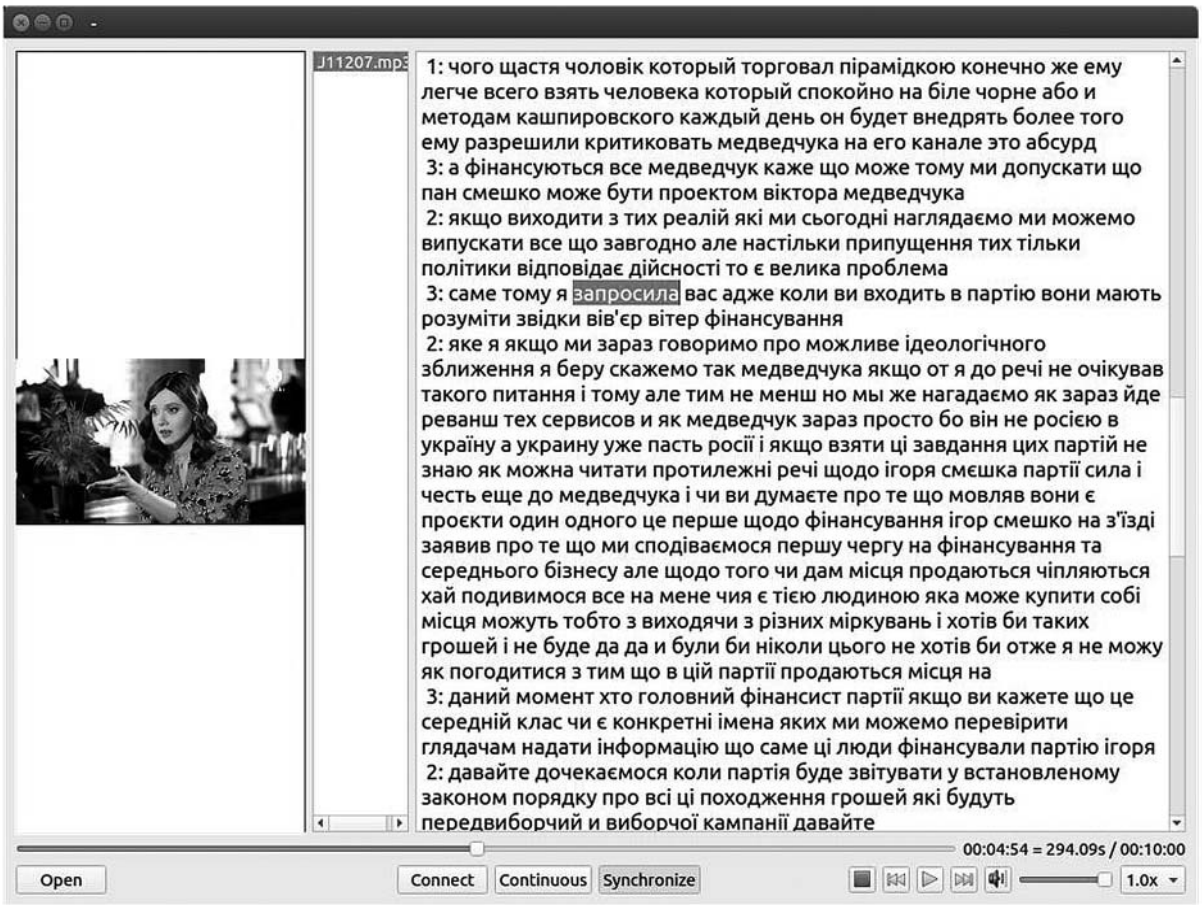

Fig. 6. Broadcast transcribing user interface

speaker identifiers can be used to automate assignment of a person whose voice is heard in the particular segment.

The *i*-vector approach used in feature space for SAT is also applicable for speaker diarization [10]. Thus we extract *i*-vectors for each segmentation, representing different speaker traits, in accordance to the detected speech activity (Fig. 3). Then the agglomerative cluster procedure is performed by merging to closest clusters, re-segmenting, updating the similarity matrix between cluster pairs and iterating the mentioned steps until the speaker diarization system reaches the stopping criterion and provides the final segmentation. Fig. 4 illustrates the diarization result for two hypothetisized speakers.

Experimental research showed diarization error rate (DER) about 12% for broadcast segments containing up to 5 speakers. This DER figure, is effective for speaker trait automation, however, DER degrades with growing the number of speakers, presence of short speech segments (less than 3 seconds), noise, distortions and speaker overlaps.

The stopping criterion for speaker diarization system is balanced in order to have a slightly larger number of hypothesized rather than reference speakers so, when editing speaker attributes, transcriptionist mostly unites speaker identity hypotheses, which obviously takes less efforts comparing to splitting hypothetical speaker identities.

The verified speaker attributes might be used to train individual speaker models based on *i*-vector. Once sufficient segments are accumulated for the speaker, a respective individual *i*-vector is extracted and written to the speaker file that is associated

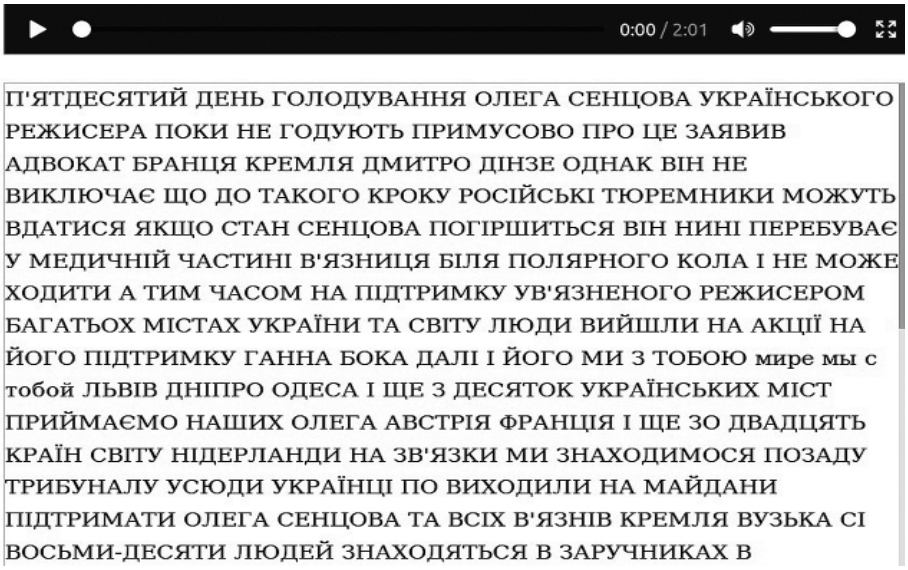

Fig. 7. Result of broadcast speech file conversion to the raw text

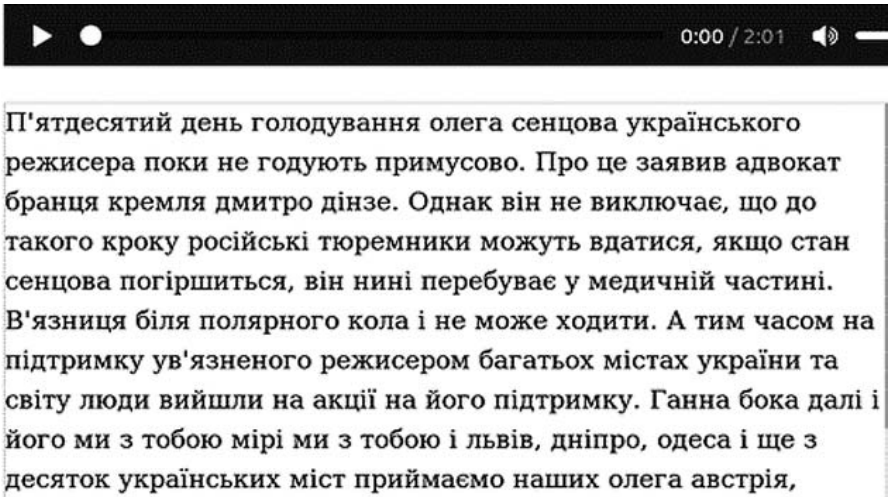

Fig. ***8.*** Automatically restored punctuation

with the specific person. Then the most similar individual i-vector to the current segment *i*-vector is used to identify a speaker as shown in Fig. 5.

Fig. 6 illustrates an experimental graphical user interface for broadcast transcribing automated with results of speech-to-text conversion and speaker individuality modeling. Text and speech signal are synchronized so that a user can simultaneously listen and look through an audiovisual segment and correct the related text and speaker attributes.

Speech-to-text conversion system output consists of words contained in the working recognition dictionary.

Thus, covering the numbers between 0 and 1000000 would mean that one million of words must be introduced to the vocabulary not counting all inflections. Since all valid numbers cannot be represented by any dictionary, we consider extraction numbers by word sequences consisting of certain basic numeric words. We use a rule-based sequence-to-sequence conversion procedure that initially has been applied to bi-directional grapheme-to-phoneme conversion [12]. The introduced multilingual rules allow for word sequence segmenting into numeric and generic words and extract digital spelling for any integer number.

For punctuation restoration, we apply a model based on Recurrent Neural Nets [11]. According to this model, the encoder sequentially encodes word sequences optionally accomplished with prosody and the decoder decodes the text with restored punctuation. The model parameters were estimated on a Ukrainian text corpus containing about 10 million sentences extracted from news web-sites so no prosody features were used.

The model is evaluated in terms of over-all precision, recall and F1-score. During testing, the punctuation with highest probability according to the model output is chosen. The development (dev) and test sets consist of 25000 and 32000 words respectively. The hyperparameters giving the best result for the development set were used to evaluate the test set.

Table 1. **Punctuation restoration results**

Set	Precision, %	Recall, %	F1-score, %
dev	82,3	59,6	68,1
test	84,6	53,9	65,9

The developed web-interface allows for listening the speech signal synchroneously with the recognized text and making corrections to words as shown in Fig. 7. The Fig. 8 illustrates how the restored punctuation facilitates perception of the extracted text.

Ukrainian broadcast episodes sometimes contain multilingual speech. Speakers may alter the language spontaneously articulating lexical and phonetical fusions. In this work we tried to cope language changing issues without prior language detection by introducing merging of basic phoneme alphabet, vocabulary and text corpora in order to create unified acoustical, pronunciation and language models. The similar approach was used e.g. in the automatic speech recognition research for several East Slavic Languages [16].

Fig. 9 illustrates the proposed basic phoneme alphabet that includes the garbage phoneme SIL, followed by stressed and non-stressed, total 10, vowels and ordinary (hard, marked with stroke) and palatalized (soften, marked with stroke), total 37, consonants.

A web-interface presented in Fig. 10 illustrates the result of text-to-speech conversion for a bilingual segment where words belonging lexically to different languages are indicated with the character case.

SIL а а’ е е’ и и’ і і’ о о’ у у’ б б’ в в’ г г’ ґ ґ’ д д’ ж ж’ з з’ й к к’ л л’ м м’ н н’ п п’ р р’ с с’ т т’

Fig. 9. Merded basic phoneme alphabet

Results

To assess speech-to-text results, firstly, for the developer set, the least error-prone hyper-para-meters in terms of word error rate (WER) are being searched and fixed then WER is estimated for the test set.

Two target languages, Ukrainian and Russian, were considered as covering the most of Ukrainian broadcast. Model parameters were estimated in single and merged language modes by respective speech and language testing sets. Proceeding from about 500 hours of the over-all speech corpus, approximately 230 hour training sets and 10 hour development and training sets were selected for each language. Training sets for language model were extracted mostly from news web-sites so we received about 2GB of text data per language.

Table 2. **Speech-to-text experimental results**

Target Language	LM order	Model’s language mode	Dev %WER	Test %WER
Ukrainian	3	Single	12,6	11,9
	3	Merged	13,0	12,6
	4	Single	12,5	12,0
	4	Merged	13,0	12,5
Russian	3	Single	14,5	16,0
	3	Merged	15,6	20,6
	4	Single	14,6	16,2
	4	Merged	15,7	20,7

The results presented in Table 2 show that lexical context wider than a 3-gram LM can provide gave no improvement. Ukrainian language looks less sensitive to model language merging. Still more attention should be paid to the selection of development and test sets, since, for the considered languages, test set WER figures have opposite dynamics comparing to development sets.

614.43
море и а пиратства интересуют извините всех пиратства
интересует и испанцев в пиратстве интересует и англичан
пиратство интересует и немцев это общие гуманитарные вопросы
понимаете это вопросы международной политики мы перешли
перевезли наш конфликт который россия хотела сделать
локальным в азове на конфликт общий общечеловеческий
александр ЗРОЗУМІЛО ЩО ВИ ПАНЕ ОЛЕЖЕ ВИ
ПРОКОМЕНТУЄТЕ АЛЕ БУКВАЛЬНО ЗА КІЛЬКА ХВИЛИН ТОМУ
ЩО ЗАРАЗ ДО НАС ДОЛУЧАЮТЬСЯ МІСТО ВЕДУЧИЙ ТАРАС
БЕРЕЗОВЕЦЬ ВІН ЗНАХОДИТЬСЯ У ВІЛЬНЮСІ ТАРАС КАЖЕ
МАТЕРІ ДОБРОГО ВЕЧОРА А ТАК Я НЕ ПОЧУЛА ТАРАСА АЛЕ
МЕНІ ЗДАЄТЬСЯ ЩО ТАРАС ЧУБАЙ МЕНЕ ДОБРОГО ВЕЧОРА
ДОБРОГО ВЕЧОРА РОЗКАЖИ ТИ РАЗ І ДЕ ЗНАХОДЯТЬСЯ

Fig. 10. Bilingual speech-to-text conversion

Comparing WERs for models trained on one (single language mode) and two (merged language mode) languages allows for estimating the language error rate (LER) of the system proceeding from an assumption that the WER degradation is introduced by the incorrect language recognition. Thus, LER for Ukrainian is about 5% on average.

Conclusions

This work presents such a level of speech signal recognition that allows for broadcast transcription process accelerating and effective search in huge amounts of spoken information.

The implemented model of multilingual speech recognition does not require a preliminary recognition of the language and has some potential to model the lexicon complementarity.

The implemented scheme speech-to-text conversion made it possible to obtain the result of the broadcast recognition in a convenient appearance for perception and modifying by a human as well as for subsequent automatic processing. Particularly, according to the acquired text, topics can be extracted, actual meanings can be tracked (proper names, numbers, dates, etc.), punctuation marks improves the perceptual experience of the text and, in general, the cost of manual editing to get the final transcript can be reduced.

Future research includes:

- punctuation restoration in dependence of prosody and multilingual modeling;
- speaker identification implementation;
- episode segmentation;
- tonality and toxity detection by speech;
- recognition of specified noise types (laugh, applause etc.);
- overall text reconstruction.

It is also planned to adapt the existing speech-to-text conversion system for online speech processing, which make the system capable for text dictating, subtitles generation and modeling a spoken dialogue between humans and technical systems.

REFERENCES

1. Vintsiuk, K., 1987. Analysis, recognition and interpretation of speech signals. Kyiv: Naukova dumka, 264 p.
2. Furui, S., 2005. "50 years of progress in speech and speaker recognition". In Proc. of 10th Int. Conf. "Speech and Computer", Patras, Greece, pp. 1–9.
3. Hinton, G., Deng, L., Yu, D., Dahl, G. et al., 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition. Signal Processing Magazine, IEEE, vol. 29, no. 6, pp. 82–97. https://doi.org/10.1109/MSP.2012.2205597
4. Mohri, M., Pereira, F.Riley, M., 2006. "Speech recognition with weighted finite-state transducers". Springer Handbook on Speech Processing and Speech Communication. Springer-Verlag, pp. 559–584. https://doi.org/10.1007/978-3-540-49127-9_28
5. Allauzen, C., Riley, M., Schalkwyk, J., Skut, W. Mohri, M., 2007. OpenFst: A General and Efficient Weighted Finite-State Transducer Library. In Proc. CIAA.
6. Sazhok, N.N., Robeiko, V.V., Fedoryn, D.Ya., Seliukh, R.A., 2015. "Broadcast Speech-to-Text System for the Ukrainian". Upravluusie sistemy i masiny, 6, pp. 66–73. [Сажок Н. Н., Робейко В.В., Федорин Д.Я., Селюх Р.А. Система преобразования телерадиовещания в текст для украинского языка. УСиМ, 2015, № 6. С. 66—73]. (In Russian).
7. Sazhok, M.M., Marikovskyy, O.V., Martynenko, M.R., Robeiko, V.V., Seliukh, R.A., Fedoryn, D.YA., 2016. "Systema avtomatychnoho monitorynhu mediynoho prostoru na osnovi tekhnolohiy rozpiznavannya slukhovykh i zorovykh obraziv". Intelektualni systemy pryynyattya rishen ta problemy obchyslyuvalnoho intelektu: Materialy mizhnarodnoyi naukovoyi konferentsiyi. Zaliznyy Port, pp. 309–310. [Сажок М.М., Маріковський О.В., Мартиненко М.Р., Робейко В.В., Селюх Р.А., Федорин Д.Я. Система автоматичного моніторингу медійного простору на основі технологій розпізнавання слухових і зорових образів. Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту: Матеріали міжнародної наукової конференції. Залізний Порт, 2016. С. 309-310.]. (In Ukrainian).
8. Zheng-Hua Tan, Achintya kr. Sarkar and Najim Dehak, "rVAD: An Unsupervised Segment-Based Robust Voice Activity Detection Method," Computer Speech and Language, 2019.
9. Dehak, T., Kenny, P., Dehak,, Dumouchel, P., Ouellet,P., 2011. "Front-End Factor Analysis for Speaker Verification", in IEEE Transactions on Audio, Speech, and Language Processing, 19(4), pp 788–798. https://doi.org/10.1109/TASL.2010.2064307.
10. Zewoudie, A.W., Luque,J., Hernando, J., 2018. "The use of long-term features for GMM- and i-vector-based speaker diarization systems". EURASIP Journal on Audio, Speech, and Music Processing, 14. https://doi.org/10.1186/s13636-018-0140-x
11. Tilk, O., Alumae, T., 2016. Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration. Interspeech, pp. 3047–3051. https://doi.org/10.21437/Interspeech.2016-1517
12. Robeiko, V.V., Sazhok, M.M., 2011. "Bahatoznachna bahatorivneva model peretvorennya orfohrafichnoho tekstu na fonemnyy". Shtuchnyy intelekt, 4. Donetsk, pp. 117–125. [Робейко В.В., Сажок М.М. Багатозначна багаторівнева модель перетворення орфографічного тексту на фонемний. Штучний інтелект. № 4'2011. Донецьк, 2011. С. 117-125]. (In Ukrainian).
13. Povey, D. "The Kaldi Speech Recognition Toolkit", Povey D., Ghoshal A., Boulianne G. al, IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.
14. CMU Dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict/
15. Shyrokov, V.A., Manako, V.V., 2001. "Orhanizatsiya resursiv natsionalnoyi slovnykovoyi bazy. Movoznavstvo", pp. 3–13. [Широков В.А., Манако В.В. Організація ресурсів національної словникової бази. Мовознавство. №5. 2001. С. 3–13.]. (In Ukrainian).
16. Safarik, R., Nouza, J., 2017. "Unified Approach to Development of ASR Systems for East Slavic Languages". In: Camelin N., Esteve Y., Martin-Vide C. (eds) Statistical Language and Speech Processing. SLSP 2017. Lecture Notes in Computer Science, vol 10583. Springer, Cham. https://doi.org/10.1007/978-3-319-68456-7_16

Received 26.11.2019

М.М. Сажок, канд. техн. наук, зав. відділом, Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України, просп. Академіка Глушкова, 40, Київ, 03187, Україна, sazhok@gmail.com,
Р.А. Селюх, м. наук. співробітник, Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України, просп. Академіка Глушкова, 40, Київ, 03187, Україна, vxm112@gmail.com
*Д.Я. Федори*н, м. наук. співробітник, Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України, просп. Академіка Глушкова, 40, Київ, 03187, Україна, dmytro.fedoryn@gmail.com
В.В. Робейко, науковий співробітник, Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України, просп. Академіка Глушкова, 40, Київ, 03187, Україна, valya.robeiko@gmail.com
О.А.Юхименко, науковий співробітник, Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України, просп. Академіка Глушкова, 40, Київ, 03187, Україна, enomaj@gmail.com

ЗАСОБИ РОЗПІЗНАВАННЯ МОВЛЕННЄВОГО СИГНАЛУ ДЛЯ ОЦИФРОВУВАННЯ УКРАЇНСЬКОГО МЕДІЙНОГО ПРОСТОРУ

Вступ. Застосовуючи розпізнавання мовлення для оцифровування медійного простору, ми розглядаємо мовленнєвий сигнал, отриманий у різних акустичних умовах від осіб, що не лише мають індивідуальні особливості вимови, а й розмовляють різними мовами. Отже, перетворення мовлення на текст має бути інваріантним до широкого класу шумів і завад, а також спотворень, які вносяться під час стискання мовленнєвого сигналу. Налаштування системи має відбуватися не лише на акустичні особливості диктора, а й на мову, якою розмовляє та чи інша особа, в тому числі, здійснюючи перехід з однієї мови на іншу й у зворотному напрямку.

Методи. При автоматичному перетворенні на текст застосовується метод, основні складові якого засновані на підходах генеративної моделі (*HMM*), апроксимації областей спостереження сигналу з використанням сумішей нормального закону (*GMM*) та покращення якості цієї апроксимації засобами глибокого навчання (*DNN*). Для моделювання акустичних особливостей людини застосовується підхід *i*-vector, що також дає змогу визначати моменти зміни мовця. Скінченні автомати та рекурентні нейромережі застосовано для поліпшення сприйняття тексту людиною та для подальшого його автоматичного оброблення. Злиття моделей двох мов дало змогу ефективно обробляти спонтанне перемикання з однієї мови на іншу.

Результати та висновки. Реалізована схема перетворення мовлення на текст дала змогу отримати результат розпізнавання фонограм телерадіомовлення у формі, зручній і для користувача-людини, і для подальшої автоматичної обробки. А саме, за отриманим текстом зрозуміло, про що йдеться, відстежується фактичний матеріал (власні назви, числа, дати тощо), розділові знаки полегшують сприйняття тексту, і загалом зменшуються затрати на ручне редагування для отримання кінцевої стенограми.

Ключові слова: *мовлення, мовленнєвий сигнал, аналіз, розпізнавання, розуміння, синтез.*

Н.Н. Сажок, канд. тех. наук, зав. отделом, Международный научно-учебный центр информационных технологий и систем НАН и МОН Украины, просп. Академика Глушкова, 40, Киев, 03187, Украина, sazhok@gmail.com,

Р.А. Селюх, м. научн. сотрудник, Международный научно-учебный центр информационных технологий и систем НАН и МОН Украины, просп. Академика Глушкова, 40, Киев, 03187, Украина, vxm112@gmail.com

Д.Я. Федорин, м. научн. сотрудник, Международный научно-учебный центр информационных технологий и систем НАН и МОН Украины, просп. Академика Глушкова, 40, Киев, 03187, Украина, dmytro.fedoryn@gmail.com

В.В. Робейко, научный сотрудник, Международный научно-учебный центр информационных технологий и систем НАН и МОН Украины, просп. Академика Глушкова, 40, Киев, 03187, Украина, valya.robeiko@gmail.com

А.А. Юхименко, научный сотрудник, Международный научно-учебный центр информационных технологий и систем НАН и МОН Украины, просп. Академика Глушкова, 40, Киев, 03187, Украина, enomaj@gmail.com

СРЕДСТВА РАСПОЗНАВАНИЯ РЕЧЕВЫХ СИГНАЛОВ ДЛЯ ОЦИФРОВКИ УКРАИНСКОГО МЕДИЙНОГО ПРОСТРАНСТВА

Введение. При применении распознавания речи для оцифровки медийного пространства рассматривается речевой сигнал, полученный в различных акустических условиях от лиц, не только имеющих свои особенности произношения, но и на разных языках. Итак, преобразование речи на текст должно быть инвариантным к широкому классу шумов и помех, а также искажений, вносимых при сжатии речевого сигнала. Настройка системы должна проводиться не только на акустические особенности диктора, но и на язык, на котором говорит то или иное лицо, в том числе, осуществляя переход с одного языка на другой и наоборот.

Методы. При автоматическом преобразовании в текст применяется метод, основные составляющие которого основаны на подходах генеративной модели (*HMM*), аппроксимации областей наблюдения сигнала с использованием смесей нормального закона (*GMM*) и улучшения качества этой аппроксимации средствами глубокого обучения (*DNN*). Для моделирования акустических особенностей человека применяется подход *i*-vector, что также позволяет определять моменты смены говорящего. Конечные автоматы и рекуррентные нейросети примененеы для улучшения восприятия текста человеком и для дальнейшей его автоматической обработки. Слияние моделей двух языков позволило эффективно обрабатывать спонтанное переключение с одного языка на другой.

Результаты и выводы. Реализованная схема преобразования речи в текст дала возможность получить результат распознавания фонограмм телерадиообщения в удобном виде, как для пользователя-человека, так и для дальнейшей автоматической обработки. А именно: по полученному тексту понятно, о чем идет речь, отслеживается фактический материал (собственные названия, числа, даты и т.д.), разделительные знаки облегчают восприятие текста, и вообще уменьшаются затраты на ручное редактирование для получения конечной стенограммы.

Ключевые слова: *речь, речевой сигнал, анализ, распознавание, понимание, синтез.*