THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 20, Number 2/2019, pp. 189–196

CROSS-LINGUAL SPEECH RECOGNITION BETWEEN LANGUAGES FROM THE SAME LANGUAGE FAMILY

Andrej ZGANK[*]

[*] University of Maribor, Faculty of Electrical Engineering and Computer Science,
Institute of Electronics and Telecommunications, Maribor, Slovenia
Corresponding author: Andrej ZGANK, E-mail: andrej.zgank@um.si

Abstract. This paper presents a detailed analysis of how languages, members of different language subfamilies, influence word accuracy in the case of cross-lingual speech recognition. Cross-lingual speech recognition presents an efficient method for developing an automatic speech recognition system for ICT applications and services in the case of under-resourced languages, where the majority of Eastern and Central European languages can be classified. In the Romance language family experiments, Spanish was selected as the source and Romanian as the target language. The experiments were carried out using the Spanish SpeechDat(II) speech database and a dedicated Romanian speech database. The Spanish source automatic speech recognition system was trained using the MASPER system. A cross-lingual phoneme mapping approach based on expert knowledge was proposed for the Romanian-Spanish language pair. The Romanian cross-lingual speech recognition system achieved word accuracy of 80.48%.

***Key words*:** automatic speech recognition, cross-lingual phoneme mapping, expert-driven approach, similar languages.

1. INTRODUCTION

Automatic Speech Recognition (ASR) is one of the human-computer interfaces which ensures natural communication and a high level of quality of experience [1]. The prerequisite for building automatic speech recognition systems is the availability of spoken language resources. The main obstacle for their availability is the high production cost and the long time needed for annotation and transcription work. As a consequence, spoken language resources are available for major world languages like English, Mandarin, Spanish, German, but a high proportion of the 6,000 world languages belongs to the group of under-resourced languages. The majority of Eastern and Central European languages can be classified into the under-resourced category, which presents a challenge for the development of state-of-the-art ICT technology nowadays used in mobile applications and services, intelligent ambiance and cloud services. An important additional advantage of cross-lingual automatic speech recognition is that it enables quick development of solutions for a new language. This can be beneficial in the case of mobile applications, where the short development time presents an important factor.

The possible solution to bypass the problem of non-existing language resources is to use cross-lingual speech recognition. In this case, a system for under-resourced target language was developed using the available system/resources from the source language. The transfer from source to target language can be carried out using two approaches [2]. The first one is based on expert knowledge, where acoustic-phonetic properties of languages are the starting point to find the most similar phoneme pairs in both languages [3,4]. The second approach is based on a data-driven metric, which is deployed to estimate the similarity between the source and target phonemes [5]. Various authors have used these approaches, producing high-quality cross-lingual speech recognition results [2].

Research on cross-lingual speech recognition of some Eastern and Central European languages [6,7] was also partly done in the MASPER initiative [8-10]. The languages included were Slovak [11], Hungarian and Slovenian, with the addition of German and Spanish. One of the results of the MASPER cross-lingual

speech recognition was the observation that the similarity between Slavic languages, reflected in the language family/subfamilies, plays an important role. Slovak, belonging to the Western Slavic subfamily, produced the best results for the Slovenian target language, which belongs to the Western group of the South Slavic subfamily. On the other hand, the combination of the Spanish source language with the Slovenian target language showed low automatic speech recognition accuracy. The open research question is how language similarity and language subfamily relations influence automatic speech recognition accuracy in the case of Romance languages, especially in the case when the source and target languages were influenced by different factors in the past. The results of such an analysis could be beneficial for future cross-lingual speech recognition, as it could ease the selection of optimal language pairs for different language families. This paper presents a detailed analysis of cross-lingual speech recognition for a Romance language pair, where Spanish and Romanian were chosen as the optimal languages to conduct the experiments. As the Spanish language has widely available speech databases, it was selected as the source language, and Romanian, as an under-resourced language, was the target language. The paper proposes a new Spanish-Romanian phoneme mapping approach based on expert knowledge, which was used to complete the cross-lingual speech recognition experiments.

The paper is organized as follows. Section 2 presents the cross-lingual speech recognition. The spoken language resources are described in Section 3. The experimental setup is presented in Section 4. The automatic speech recognition results are described in Section 5, while the conclusions are given in Section 6.

2. CROSS-LINGUAL SPEECH RECOGNITION

Cross-lingual automatic speech recognition presents an important approach on how to develop a system for under-resourced languages in an efficient way. The basic idea is to use available resources (speech database, acoustic models, and lexicon) from the source language, and transform them into a form appropriate for automatic speech recognition in an unseen target language. The basic principle is presented in the form of a block scheme in Fig. 1.

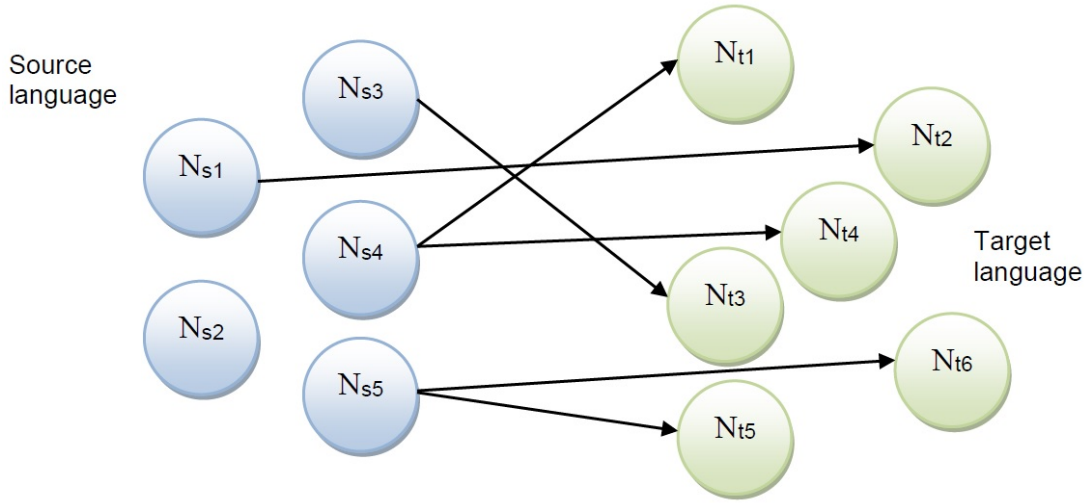

Fig. 1 – Cross-lingual speech recognition scheme.

The source language phonemes ϕ_s are transferred into a target language ϕ_t according to some metric. The mapping can be one to one or one to many, as long as all target language phonemes are covered by a similar source phoneme. The first cross-lingual speech recognition systems were presented in the ’90s by different authors [2]. Although speech recognition systems for major languages have improved by several degrees over the last two decades, the 6,000 languages used actively in the world still present an important topic for cross-lingual speech recognition.

One of the key issues for cross-lingual speech recognition remains the question of similarity between the source and target language. This question becomes even more important in the case when multilingual

acoustic models are used as the source [2]. Optimal selection of the source language improves the accuracy and performance of a cross-lingual speech recognition system significantly in such a case.

In the case of this experiment, the source language was Spanish and the target language was Romanian. Both languages belong to the Romance language family, but to different subfamilies. According to the linguistic theory, Spanish belongs to the subfamily of Western Romance languages, and Romanian to the groups of Eastern Romance languages. Both subfamilies were subjected to various influences in the past, which are nowadays reflected in phonetics, semantics and syntax differences between the two language families. These differences are much more present in this case than in the case of languages which belong to the same subfamily.

The language phonetics play a key role in cross-lingual speech recognition. The Spanish source phoneme set has 31 elements, and the Romanian has 30 elements. Although the phoneme set size is similar for both languages, there are still significant differences present, especially in the case of vowels and some other particular phonemes. Vowels have an important influence on automatic speech recognition accuracy. Using the expert knowledge about acoustic-phonetic properties and phoneme occurrences relevant from acoustic models` training point of view, mapping was carried out from the target phonemes to the existing source language phonemes. The proposed phoneme mapping from Romanian to Spanish is presented in Table 1. All phonemes are transcribed using the SAMPA nomenclature.

Table 1

Cross-lingual phoneme mapping from Romanian to the Spanish language

mapping																
Romanian	a	@	1	e	i	i_0	o	u	e_X	j	o_X	b	d	g	k	p
Spanish	a	e	o	e	i	i	o	u	e a	j	u a	b	d	g	k	p
Romanian	t	m	n	l	f	v	s	S	z	h	r	Z	ts	tS	dZ	
Spanish	t	m	n	l	f	B	s	s	z	x	r	z	tS	tS	d Z	

The 20 Romanian phonemes were mapped into their Spanish equivalents, according to acoustic-phonetic characteristics. The remaining 11 Romanian phonemes were without direct Spanish equivalent, therefore, the most similar candidates were selected.

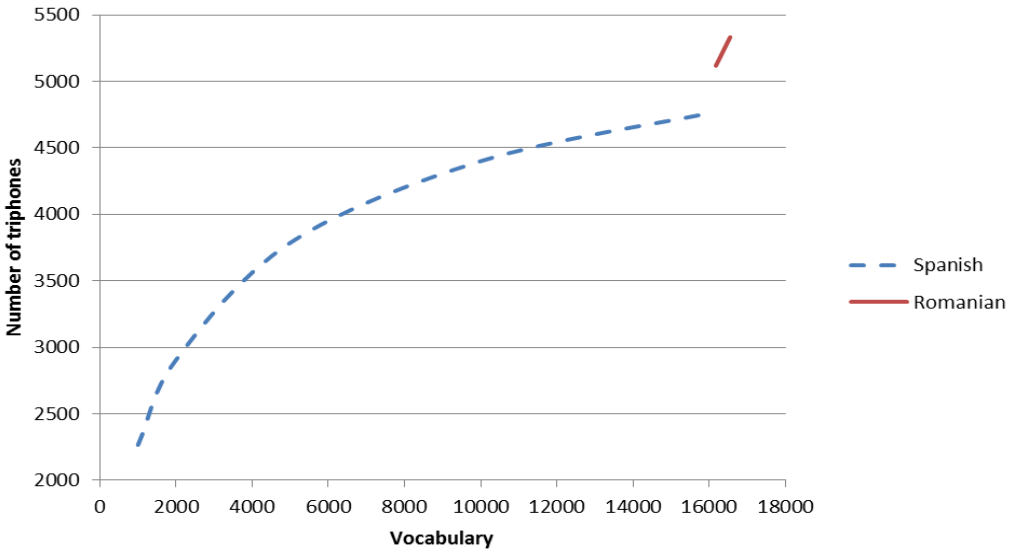

Fig. 2 – Number of new triphones.
The addition of 375 Romanian words increased the number of new triphones further to 5,333 – see marker.

Because of differences in phonetic structure, the mapped cross-lingual acoustic models introduced new unseen triphones, which have to be processed using the phonetic decision tree based clustering. The graph in Fig. 2 presents the dependency between the size of the phonetic vocabulary and number of triphones. The graph shows the increase in the number of triphones for Spanish, where the number of new triphones

increases from 2,265 for 1,000 words to 4,759 for 15,826 words. Saturation can be observed of the number of new triphones. When 375 Romanian words were added, the number of new triphones increased significantly to 5,118. Additionally, the added 375 Romanian words increased the number of new triphones further to 5,333. This analysis shows that, from the acoustic modeling point of view, both languages have different phonetic structure.

3. SPEECH DATABASES

The under-resourced languages are a typical focus of cross-lingual speech recognition. Because pattern recognition approaches are usually applied, the selection of an appropriate speech database for the source language has an important influence. The goal of the presented work was to study and analyze the acoustic influence between languages belonging to the same language family, which is reflected in the cross-lingual speech recognition accuracy. This resulted in a decision to use isolated words for the speech recognition scenario, as the influence of language modeling was omitted in this case.

3.1. Spanish 4000 FDB SpeechDat(II) speech database

The Spanish 4000 FDB SpeechDat(II) speech database [12] was used for source acoustic modeling. This spoken language resource belongs to a large family of SpeechDat databases (i.e. SpeechDat(II), SpeechDat(E), SpeechDat(M), SpeechDat Car) which cover more than 40 languages. All SpeechDat databases share the same design guidelines [13], and are used widely in different automatic speech recognition experiments and applications. The majority of SpeechDat databases is available by ELRA/ELDA. The SpeechDat(II) category of databases was recorded over the fixed telephone from a real-world environment, usually in a home or office environment. The speech signal was sampled with 8 kHz and is coded with the aLaw speech codec.

The subset with the first 1,000 speakers was selected from the Spanish database for the experiments. This was necessary to be able to compare the achieved results with other languages [9] which have a smaller size of language resources available. Each speaker in the database uttered 42 different phrases or sentences, including continuous digits, city names, phonetically balanced words, and sentences. Phonetically balanced material is particularly important for cross-lingual speech recognition, as it guarantees the correct acoustic representation of language. The uttered speech was transcribed manually, including the phonetic lexicon with 15,826 entries based on the SAMPA phoneme set. In the case of the Spanish SpeechDat(II) database, the phonetic set has 31 different phonemes, which can be used for source acoustic modeling. In addition to phonemes, four additional acoustic effects were also annotated in the database transcriptions: speaker noises, background noises, breaths and onomatopoeias. Only speaker noises and breaths were used in our case for acoustic modeling, as in the case of other two annotations, the exact time frame isn't specified, which makes acoustic modeling unpredictable.

The first 800 speakers were included in the training set, while the remaining 200 were selected for the test set, which was needed to evaluate the source automatic speech recognition accuracy. The evaluation was carried out with the following three test scenarios:

- A-set: application words, used for command and control in intelligent ambiance solutions,
- O-set: city names, used for mobile application and services,
- W-set: phonetically balanced words.

Similar test sets were used frequently in other cross-lingual speech recognition experiments [9, 10, 11], which guarantee the possibility to compare the results. The characteristics of the source language test sets are presented in Table 2.

The Table 2 with characteristics shows the different complexities of three test sets. The simplest one, with the smallest number of words in the vocabulary, is the test set A. Although it has short words, which are sometimes difficult to be separated by an automatic speech recognition system due to the end phoneme reductions, it is still anticipated that this test set produces the highest speech recognition accuracy. City names (O) is a test set of medium complexity with 749 words in the dictionary. Its main particularity lies in the acoustic diversity of included words. A drawback of this test set is that some of the foreign city names

were pronounced with a non-native accent. The most complex test set is the W, with phonetically balanced words, which has 2,987 entries in the vocabulary. The longer, phonetically balanced words, included in this test set, are usually well distinguishable by the automatic speech recognition system, which partially reduces the overall complexity of the W test set. Nevertheless, it is still anticipated that the lowest speech recognition accuracy will be produced with it.

Table 2

Source language (Spanish) test set characteristics.

Spanish test set	Number of words	Average length of word
A	64	6.34
O	749	6.11
W	2,987	7.75

3.2. Romanian speech database

A dedicated speech database was built for Romanian, which was used as the target language. The main source of Romanian speech recordings was the project SRoL "Voiced Sounds of Romanian Language" [14], where the recordings were acquired in neutral emotion. The source of the remaining speech recordings was various on-line available materials. Recordings of isolated words and short phrases were selected for the Romanian evaluation set. This type of speech material has priority, as it is the most suitable for the evaluation of acoustic models. The Romanian evaluation set comprised 37 speakers, who uttered the 420 different utterances used for cross-lingual speech recognition.

The initial speech recordings have different characteristics, where the SRoL material was recorded with 22.05 kHz 16-bit mono sampling, and the remaining using the AAC with 192 kbps, 44.1 kHz. The speech recordings were downsampled to 8 kHz aLaw format, to be compatible with the format of speech recordings in the Spanish SpeechDat(II) database. The segmentation of recordings to a form appropriate for automatic speech recognition was done in an automatic way, using the GMM voice activity detection approach. The word level transcriptions, needed for evaluating the speech recognition results, were created from acquired transcriptions and closed captions.

The Romanian phonetic transcriptions were generated with the NaviRO phonetic dictionary [15], and its phoneme set, with 30 elements, was used for the Romanian cross-lingual speech recognition. The Romanian cross-lingual speech recognition evaluation scenario contained a grammar with 750 isolated words and short phrases. Its complexity is comparable with the Spanish speech recognition scenario O, which was used for evaluation of Spanish source acoustic models. This similarity is essential if the comparison of speech recognition accuracy between languages belonging to the same family would be performed.

4. EXPERIMENTAL SETUP

The speech recognition experimental setup is based on the MASPER system [8], which is one of the frequently used approaches to conduct cross-lingual or multilingual speech recognition. The baseline presents continuous density Hidden Markov Models, with Gaussian probability density function (1), defined as:

$$b_j(\mathbf{o}_t) = \frac{1}{\sqrt{(2\pi)^n \left|\Sigma_j\right|}} e^{-\frac{1}{2}(\mathbf{o}_t - \mu_j)\Sigma_j^{-1}(\mathbf{o}_t - \mu_j)} \tag{1}$$

Three state left-right topology is used for the acoustic modeling, which was carried out for the source language. Acoustic models produced for the experimental setup and cross-lingual speech recognition are speaker independent.

The prerequisite for an acoustic training procedure is feature extraction, which transforms the input acoustic signal into the data needed for pattern recognition. The speech signal is windowed using a 25 ms

Hamming window, which is shifted each 10 ms. The resulting frames are pre-emphasized, and then 12 Mel-cepstral coefficients and energy are calculated. These 13 features are used for calculating the first and second order derivatives, which model the stable phase of the speech signal. In the final step, the cepstral mean normalization is performed, which reduces the influence of the recording channel on the speech signal. This normalization process is important for cross-lingual speech recognition, as usually more than one speech database is needed. In a typical setup, each speech database has at least slightly different recording conditions. The final feature vector has 39 different coefficients.

The first step in the training source acoustic models (2) is devoted to context-independent monophone models, where the task is to estimate the model parameters:

$$\lambda = \{A, B, \pi\} . \tag{2}$$

The acoustic models' parameters are initialized with global values estimated on the randomly selected subset of the full training set. A few iterations of Baum-Welch reestimation are applied to train the acoustic models.

The trained acoustic models are used for improving the speech transcriptions using the forced-realigning procedure. The spoken language resources are collections of big data, and are always imperfect. The forced-realigning procedure excludes from the training set those utterances with degraded acoustic conditions (low SNR, echo, channel distortions), or speech (strong accent, hesitations, fragments). In addition, transcriptions with error can be handled in such a way. Approximately 0.19% of the training set was excluded as outliers during the first application of the forced-realigning procedure.

The improved speech transcriptions were used in the second step of acoustic training, were context-independent models were initialized again from scratch. After a few iterations of Baum-Welch reestimation, the number of Gaussian probability density functions was increased in a step-wise manner from 1 to 32 per state. After each increase, the acoustic models' parameters were re-estimated again. These acoustic models were then applied for the second run of the forced realignment, with the goal to improve the database transcriptions further. This time, approximately 0.05% of the training set was excluded as outliers, due to various errors or imperfections.

In the third step of acoustic modeling, the context was taken into account. The triphone acoustic models were built from the previous context-independent models. These resulted in a dramatic increase of acoustic models' free parameters which needed to be estimated. Phonetic decision tree based clustering was applied to reduce the number of free parameters, and control it in balance with the available speech material. The phonetic decision trees were induced with the Spanish broad phonetic classes defined by a human expert. The clustering was performed on a state level. A complete merger was carried out in the case where all three states of an acoustic model were clustered to the identical triphone. The phonetic decision tree based clustering resulted in 4,609 triphone acoustic models. The final acoustic models` training step was to increase the number of Gaussian probability density functions to 32 per state. This set of Spanish acoustic models was used for source language evaluation and for Romanian cross-lingual speech recognition.

5. RESULTS

The evaluation of automatic speech recognition was carried out in two parts. First, the Spanish source language speech recognizer was evaluated, to verify its suitability for the Romanian cross-lingual speech recognition, which was evaluated in the second part. Word accuracy (3) was used for the evaluation. It is defined as:

$$WA[\%] = \frac{H}{N} \cdot 100 , \tag{3}$$

where N presents the number of all words in the test set, and H the number of correctly recognized words.

The automatic speech recognition results with the Spanish source acoustic models are given in the first part. In addition to the triphone speech recognition results, the results for context-independent monophone acoustic models are also given, with the goal to evaluate the efficiency of context modeling. The three test sets A, O, and W from the SpeechDat(II) database were used for evaluation. The results in the form of word accuracy are presented in Table 3.

Table 3

Spanish source language automatic speech recognition results.

Spanish test set	monophone WA(%)	triphone WA (%)
A	92.10	97.82
O	67.21	88.52
W	60.30	71.09
average	73.20	85.81

In general, the achieved results were in the range with the values predicted in Section 3 with the database description. The highest word accuracy of 97.82% was achieved with the A test set, which was the simplest one. The O test set produced 88.52% word accuracy. The worst word accuracy of 71.09% was obtained with the W test set, which is the most complex test set. With its 2,987 different words in the dictionary, it can be classified as a middle vocabulary speech recognition scenario. The achieved results are comparable with those achieved by automatic speech recognition systems of similar complexity [9]. The difference between monophone and triphone acoustic models is higher for the more complex test scenarios (O and W), which confirms the decision to use triphone acoustic models. From these speech recognition results it can be concluded that Spanish source acoustic models present a good baseline for Romanian cross-lingual speech recognition.

The cross-lingual speech recognition results were evaluated with the Romanian test set. The Romanian word accuracy results are presented in Table 4.

Table 4

Romanian cross-lingual speech recognition results.

	monophone WA(%)	triphone WA (%)
Romanian test set	72.14	80.48

The 80.48% word accuracy was achieved with the Romanian cross-lingual acoustic models generated from the Spanish source acoustic models. The monolingual Spanish speech recognition system on a comparable test set achieved 88.52% word accuracy. The difference in speech recognition between monolingual and cross-lingual cases is approximately 8%. This result was significantly better than the original MASPER case [8,10], when Spanish and Slovenian formed the language pair. There, the word accuracy was as low as 18.04%. The achieved result confirms the hypothesis that language family similarity is one of the key factors in the case of cross-lingual speech recognition. The Romanian cross-lingual triphone acoustic models improved the word accuracy significantly, as the evaluation of the Romanian cross-lingual monophone acoustic models achieved word accuracy of 72.14%. The modeling of context in the case of cross-lingual triphones is especially important, as the triphone statistics (presented in Section 2), already indicated significant diversity in the case of phoneme structure between the Spanish and Romanian languages. The Romanian cross-lingual speech recognition results are comparable to the Romanian speech recognition results published by other authors [4,7,16,17].

The Romanian cross-lingual speech recognition results indicate the possibility of using such an approach for a real-life human-computer interface in the case of a limited domain, which would, additionally, improve the speech recognition accuracy. This could be then applied to various ICT services and applications.

6. CONCLUSIONS

The paper presented an analysis of how language family/subfamilies influence the accuracy of cross-lingual speech recognition for the language pair Spanish-Romanian. This language pair was well suited for the analysis, as it belongs to different language subfamilies, with significant differences in acoustic and phonetic characteristics, which are key factors for cross-lingual speech recognition. The additional advantage of the selected language pair was that Romanian belongs to the group of under-resourced languages, and is, as such, a typical candidate for cross-lingual speech recognition.

The results of cross-lingual mapping based on expert knowledge and the analysis of the number of triphones, confirmed the hypothesis that different language subfamilies present a challenging environment for cross-lingual speech recognition. The Romanian cross-lingual speech recognition achieved word accuracy lower by approximately 10%, which is similar to comparable cross-lingual setups with medium vocabulary size.

The results indicate that the appropriate language pairs for cross-lingual speech recognition can be selected outside the same language subfamily. The future work will be focused on carrying out the analysis for other language subfamilies, and studying this influence in the case where multilingual acoustic models will be applied as the source for cross-lingual speech recognition.

ACKNOWLEDGEMENTS

This research work was partially funded by the Slovenian Research Agency ARRS under Contract number P2-0069.

REFERENCES

1. C.H. LEE, *On automatic speech recognition at the Dawn of the 21st century*, IEICE Trans. Inf. Syst., **E86-D**, *3*, pp. 377–396, 2003.
2. L. BESACIER, E. BARNARD, A. KARPOV, T. SCHULTZ, *Automatic speech recognition for under-resourced languages: A survey*, Speech Communication, **56**, *1*, pp. 85–100, 2014.
3. T. SCHULTZ, *Multilinguale Spracherkennung – Kombination akustischer Modelle zur Portierung auf neue Sprachen*, PhD Thesis, University of Karlsruhe, Germany, 2000.
4. C. CHIVU, *Systems of continuous speech recognition for Romanian language.* Control Engineering and Applied Informatics, **7**, *4*, pp. 63–68, 2006.
5. C. NIEUWOUDT, E.C. BOTHA, *Cross-language use of acoustic information for automatic speech recognition*, Speech Communication, **38,** *1-2*, pp. 101–113, 2002.
6. G. TAMULEVICIUS, A. SERACKIS, T. SLEDEVIC, D. NAVAKAUSKAS, *Vocabulary distance matrix analysis-based reference template update technique*, Proceedings of the Romanian Academy, Series A, **16**, *1*, pp. 103–109, 2015.
7. C. CHIVU, *Applications of speech recognition for Romanian language*, Advances in Elec. and Comp. Engineering, **7**, *1*, 2007.
8. MASPER initiative, http://masper.um.si
9. A. ZGANK, Z. KACIC, F. DIEHL, K. VICSI, G. SZASZAK, J. JUHAR, S. LIHAN, *The COST 278 MASPER initiative – crosslingual speech recognition with large telephone database*, Proc. LREC 2004, Lisbon, Portugal, 2004.
10. A. ZGANK, Z. KACIC, K. VICSI, G. SZASZAK, F. DIEHL, J. JUHAR, L. LIHAN, *Crosslingual transfer of source acoustic models to two different target languages*, Proc. ROBUST 2004 Workshop, Norwich, UK, 2004.
11. J. KACUR, R. KOZICKA, R. VARGIC, *Semi-tight covariance matrices implementation in Masper HMM Training Procedure*, Proc. IWSSIP 2016, Bratislava, Slovakia, 2016.
12. H. HOEGE, H. TROPF, R. WINSKI, H. VAN DEN HEUVEL, R. HAEB-UMBACH, *European speech databases for telephone applications*, Proc. ICASSP '97, Munich, Germany, 1997.
13. H. VAN DEN HEUVEL, L. BOVES, A. MORENO, M. OMOLOGO, G. RICHARD, E. SANDERS, *Annotation in the speech-Dat projects*, International Journal of Speech Technology, **4**, *2*, pp. 127–143, 2001.
14. S.M. FERARU, H.-N. TEODORESCU, M.D. ZBANCIOC, *SRoL – Web-based resources for Llnguages and language technology e-Learning*, International Journal of Computers, Communications & Control, **V**, *3*, pp. 280–290, 2010.
15. J. DOMOKOS, O. BUZA, G. TODEREAN, *Romanian phonetic transcription dictionary for speeding up language technology development*, Language Resources and Evaluation, **49,** *2*, pp 311-325, 2015.
16. T. BOROȘ, D. TUFIȘ, *Romanian-English speech translation*, Proceedings of the Romanian Academy, Series A, **15,** *1*, pp. 68–75, 2014.
17. C.O. DUMITRU, I. GAVAT, *Progress in Speech Recognition for Romanian Language*, Advances in Robotics, Automation and Control, I-Tech, Vienna, Austria, pp. 472, 2008.

Received June 13, 2016