Online ISSN 1848-3380, Print ISSN 0005-1144
ATKAFF 57(1), 268–281(2016)

Tadej Justin, France Mihelič, Janez Žibert

Towards automatic cross-lingual acoustic modelling applied to HMM-based speech synthesis for under-resourced languages

DOI 10.7305/automatika.2016.07.1084
UDK 004.522:[004.934‘1:81‘344.2=00]

Original scientific paper

Nowadays Human Computer Interaction (HCI) can also be achieved with voice user interfaces (VUIs). To enable devices to communicate with humans by speech in the user's own language, low-cost language portability is often discussed and analysed. One of the most time-consuming parts for the language-adaptation process of VUI-capable applications is the target-language speech-data acquisition. Such data is further used in the development of VUIs subsystems, especially of speech-recognition and speech-production systems. The tempting idea to bypass a long-term process of data acquisition is considering the design and development of an automatic algorithms, which can extract the similar target-language acoustic from different language speech databases. This paper focus on the cross-lingual phoneme mapping between an under-resourced and a well-resourced language. It proposes a novel automatic phoneme-mapping technique that is adopted from the speaker-verification field. Such a phoneme mapping is further used in the development of the HMM-based speech-synthesis system for the under-resourced language. The synthesised utterances are evaluated with a subjective evaluation and compared by the expert knowledge cross-language method against to the baseline speech synthesis based just from the under-resourced data. The results reveals, that combining data from well-resourced and under-resourced language with the use of the proposed phoneme-mapping technique, can improve the quality of under-resourced language speech synthesis.

Key words: Voice user interfaces, Human language technologies, HMM-based speech synthesis, Cross-language synthesis, Under-resourced languages, UBM-MAP-GMM phoneme mapping

Primjena automatskog međujezičnog akustičnog modeliranja na HMM sintezu govora za oskudne jezične baze. U današnje vrijeme interakcija čovjeka i računala (HCI) može se ostvariti i putem govornih sučelja (VUIs). Da bi se omogućila komunikacija uređaja i korisnika putem govora na vlastitom korisnikovom jeziku, često se raspravlja i analizira o jeftinom rješenju prijevoda govora na različite jezike. Jedan od vremenski najzahtjevnijih dijelova procesa prilagodbe jezika za aplikacije koje podržavaju VUI je prikupljanje govornih podataka za ciljani jezik. Ovakvi podaci dalje se koriste za razvoj VUI podsustava, posebice za prepoznavanje i produkciju govora. Primamljiva ideja za izbjegavanje dugotrajnog postupka prikupljanja podataka jeste razmatranje sinteze i razvoja automatskih algoritama koji su sposobni izvesti slična akustična svojstva za ciljani jezik iz postojećih baza različitih jezika. Ovaj rad fokusiran je na povezivanje međujezičnih fonema između oskudnih i bogatih jezičnih baza. Predložena je nova tehnika automatskog povezivanja fonema, usvojena i prilagođena iz područja govorne autentikacije. Ovakvo povezivanje fonema kasnije se koristi za razvoj sustava za sintezu govora zasnovanom na HMM-u za manje poznate jezike. Načinjene govorne izjave ocijenjene su subjektivnim pristupom kroz usporedbu međujezičnih metoda visoke razine poznavanja jezika u odnosu na sintezu govora načinjenu iz oskudne jezične baze. Rezultati otkrivaju da kombinacija oskudne i bogate baze jezika uz primjenu predložene tehnike povezivanja fonema može unaprijediti kvalitetu sinteze govora iz oskudne jezične baze.

Ključne riječi: Govorna korisnička sučelja, tehnologije ljudskog govora, HMM sinteza govora, međujezična sinteza, oskudna jezična baza, UBM-MAP-GMM povezivanje fonema

1 INTRODUCTION

Spoken language is one of the primary means of sending or receiving information. Therefore, it is the most natural and the most common form of establishing communication between humans. One of the initial ideas, when developing the first personal computers, was to communicate with such devices also by voice. The research field of speech technologies has a long tradition from the beginning of the digital age and, finally, nowadays end users are able to use application software, currently in a limited number of languages, which is able to establish commu-

nication by voice user interfaces (VUIs) [1]. The benefits of the VUI are obvious when the operator's hands are fully occupied. Furthermore, such a communication is often the only available way when the operators are people with physical disabilities, such as blind people or people with muscular dystrophy, who use electronic devices for assistance. Today's applications for controlling devices by voice are trying to access customers in global markets and give them the opportunity to communicate in the customer's own language. Comprehensive reference work for catalogue all of the world's known living languages, *Ethnologue: Languages of the World*[1] at the time of this research reports that there are 7106 living languages. On the other hand only a small percentage offer the resources required for the implementation of Human Language Technologies (HLTs) [2].

This statement refers to the fact that obtaining such a resources is a long-term and costs-related investment. Furthermore, this also suggests that subsystems designed for VUI are strongly dependent on language resources. The recent activities in speech-related technologies point to an interest in researching language portability for multilingual applications [3] and there is also special attention devoted to under-resourced languages. As a result of political or economic interest, HLT is more attractive for some languages than others. Investing efforts in researching methods for speech generation (speech synthesis) and speech recognition as well as in machine-translations systems where such systems are designed as language independent can be remunerated not only for preserving spoken languages but also for helping humans to overcome language barriers. Researching the adaptation techniques to a specific language with the use of available speech data from other languages is one of the potential steps that could lead to low-cost language portability. When insufficient or limited language dependent speech data is available, the term under-resourced language is commonly used [4].

In this article we describe and propose one of the possible solutions for low-cost language portability involving a speech-synthesis system for under-resourced languages. We present a novel approach for finding similar phonemes based on the acoustic representations between Slovene and English for further usage in a HMM-based speech-synthesis system or even in speech-recognition systems. Based on the cross-language phone mapping table we implement a novel method for extending the parametric acoustical models estimated with the low-resourced data for HLT. For the purposes of this research we focus only on the quality of the produced speech of the proposed techniques for the speech synthesis of an under-resourced language. We report on a subjective evaluation of the synthesised speech obtained with automatic and manual phone-mapping techniques. The subjective tests reveal the improvement in the overall quality of the synthesised speech obtained with extended, under-resourced, language-dependent, acoustic models with the acoustic being modelled from a well-resourced language. Our experiments are conducted with the use of speech resources from the English CMU ARCTIC speech database [5] and a predefined small part of the Slovene speech database VNTV [6], which simulates the under-resourced language.

The article is structured as followed. First we introduce the related research work in the next subsection and describe the motivation for this research in Subsection 1.2. Section 2 describes the conditions and speech resources used in our systems implementation. It also provides a detailed description of the system development based on expert cross-language phoneme mapping as well as the automatic cross-language mapping technique. The evaluation process is described in detailed in Section 3. The experimental results of the evaluation method are presented in Section 4. We discuss the obtained results in Section 5 and end with the conclusion in Section 6.

1.1 Related work

Currently, unit selection [7] is still the underlying technique in most commercial systems. Since this requires large resources of well-recorded and labelled speech data to ensure the optimal unit coverage, its use is becoming less common for obtaining cheap and fast language portability, especially in the case of an under-resourced language. On the other hand, a statistical parametric synthesis [8] does not have such strict requirements, but also comparable quality. The parametric representation makes transformations feasible by using simple, linear, algebraic operations. This flexibility of the parametric models makes them well suited for performing adaptations or manipulations in relation to the amount of required speech data. Therefore, this is one of the most widely used, speech-synthesis techniques when synthetic speech is required in an under-resourced language.

The quality of the produced speech is actively related to the amount of speech data available during the training of the acoustic models [9]. If such data is not available, the developers are often subject to speech-data acquisition. In many cases the target-language speech data has to be recorded, transcribed and additionally labelled. Such a process is often time consuming and also requires some experiences in design. An attractive idea to overcome such a time-consuming task is to map the most similar acoustics from other languages, which already have well-established speech databases with a new language acquired for the speech synthesis.

The parametric speech synthesis framework HTS [10] is often used for implementing or developing synthesis

[1] http://www.ethnologue.com

systems with the use of cross-lingual resources. The basic approach and the most commonly used one is that of phoneme mapping, where the approximations of the most similar available sounds of the target language and the sounds in the available similar language database are found [11, 12]. Such methods require the mapping of the related phonemes in both languages. The relations between phonetic representations of different languages can be found with an ad-hoc method or with the use of different automatic approaches, which are mostly related to speech recognition. One of the most attractive ones is to build a language-dependent phoneme speech recognition system. With such a system the target language utterances (phonemes) can be recognized as the ones that are most similar to those in the other language and vice-versa. With such an established relationship the different distance measures can be calculated and are also commonly used between the HMM mixture components as in [13–15].

Although wide use of a phone mapping approach, also other solutions that are based on language adaptations are also found in literature. Developing a polyglot speech-synthesis system [16] is alternative possible solution to cross-lingual speech synthesis. It can be defined as a synthesis system, that is able to produce synthetic speech in multiple languages using the same target-speaker characteristics [17]. Commonly, such approaches are also tested with bilingual speech-synthesis systems [18]. Such an approach to multilingual speech synthesis requires the target language's data during training. If such data is available, experiments to cross-language synthesis are reported using the cross-lingual databases. Such speech databases includes the speech data, where multiple speakers, who are capable of speaking more than one language are recorded [19]. In such circumstance authors tested a state-mapping approach [20], where the main task is to find the state similarities of average voice-language-dependent acoustic models [17]. Also, frame mapping and Gaussian component mapping have been explored in [21, 22] and [23], respectively.

With the ability to influence the estimates to parameters commonly by adaptation, interpolation or even label manipulation, building speech-synthesis systems for under-resourced and also polyglot systems is one of the most attractive advantages in HMM-based speech synthesis in comparison to the unit-selection method for speech synthesis. The relatively large number of proposed methods implies a dynamic field of cross-lingual speech synthesis and also an in interest obtaining high-quality speech synthesis thought minor investments.

1.2 Motivation

As described in 1.1, the HMM-based speech-synthesis technique is well suited to the task of language adaptations and also well known for its small footprint; therefore, it is the appropriate solution for integrating not only in devices with greater capacities, but also for small embedded devices [24, 25].

Frequently, researchers are challenged with two questions in such scenarios. Firstly, what is the minimum required speech data for a target language to build a HMM-based speech-synthesis system that produces still intelligible speech and, secondly, is it possible to improve the HMM acoustic models trained with minimal speech data without collecting new additional target-language data, but just by using existing speech data from different languages and speakers?

Looking for the appropriate answers, researchers commonly encounter two sub-problems. The first one is related to different systematizations and definitions of the phoneme units between different languages. Finding the most appropriate mapping is a challenging task. The obvious method for solving such a problem is to employ a phonetic expert with a phonetic knowledge of both languages. From the developer's point of view, the most attractive solution is to find such a mapping automatically. The second problem refers to the quality of the produced speech resulting from such language mapping. The synthetic voice is commonly exposed to a strong foreign accent. The solution is often achieved with the use of HMM-based speech synthesis and the possibility of adapting the mapped acoustic models with a small amount of data available in the target language.

In this article we try to determine whether an acoustic model for a speech-synthesis system obtained by mapping similar phonemes from other well-resourced languages with the use of additional adaptations available in HMM-based speech synthesis can improve the quality of the acoustic modelling estimated from a small amount of data available in the target, under-resourced language.

The HMM-based speech-synthesis is based on context-dependent speech segments. These are related with the estimations of the parametric speech representations. Most likely, such parametrized representations are estimated with the likelihood distributions. In HMM-based acoustic modelling they are, obtained from estimations of the speech segments included in the speech database. The size of the speech database has an important role in good estimations of the parameters, and also on the actual size of different context-dependent labels. If the available speech data is rather small, the expected quantity of estimated, context-dependent, acoustical models is low. During the synthesis part, when the context-dependent label is not found between all the estimated labels in the acoustic model, the use of phonetic trees is activated to find the most similar context-dependent label. Based on the

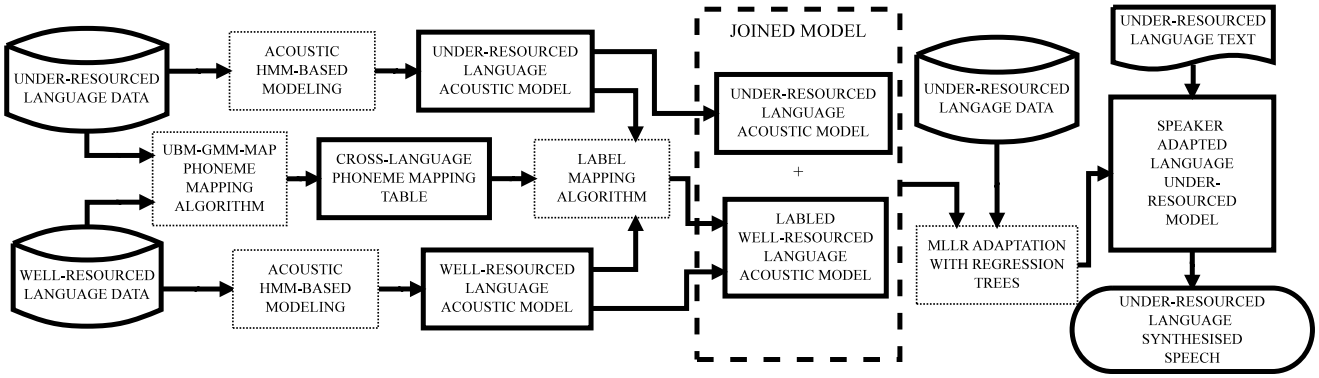

Figure 1. Overview of bilingual speech-synthesis system for under-resourced language

phonetic similarities, the synthesised speech can be produced, although the incoming text did not include the existing context-dependent label from the speech databases. Therefore, in the cases of large speech databases, a more accurate production of the speech is expected. The flexibility of HMM-based speech synthesis permits the extension or manipulation of the context-dependent speech labels. Therefore, the idea of expanding the estimated context-dependent acoustic models with the use of other language estimates and well-defined and accurate cross-language phoneme mapping with additional adaptation to the data available for the target language is a promising approach to obtain more diverse, context-dependent estimates. Because it is hard to obtain an accurate cross-language phoneme mapping and in many mapping examples only the most similar phonemes are mapped, it is hard to expect that the improvements on all speech segments in terms of produced speech quality are obtained.

In this article we investigate one of the possible procedures to obtain a synthetic voice in an under-resourced language with the use of different speakers from well-resourced speech databases. In addition, we want to experimentally, based on a subjective evaluation test, confirm the hypothesis that with the proposed method we can obtain a better quality of speech synthesis in the target, under-resourced language.

The experiments in this research investigate the Slovene-English mapping strategy. Since the English language is well covered by HLT resources, it is one of the most attractive ideas for using such speech resources to improve the performance of speech-related systems capable of producing or recognizing speech from other under-resourced languages. Because of the large language diversity between English, which is classified as a West German language related to Dutch, Frisian, German, etc. and the Slovenian language, which belongs to the South Slavic language group and is related, for example to the Croatian and Serbian languages, is hard to obtain improvements in all fields of HLT. Therefore, in this article we focus only on the acoustic label mapping of spectral models. By establishing of a cross-language phone-mapping table and by performing a simple transformation algorithm for phonetically labelled acoustic models, we also obtain larger coverage of the so-called seen acoustical models. Such joined modes of pure Slovene and English spectral acoustics are suitable for further adaptations to Slovene speech data.

2 METHODOLOGY

The experimental procedure refers to the development of an under-resourced language speech-synthesis system. The system diagram in Fig. 1 represents the basic steps for obtaining bilingual acoustic models for further implementation in the speech-synthesis system. From a well-resourced-language speech database (CMU-arctic database) and an under-resourced language database (part of the VNTV database) we define a phone-mapping table. By finding cross-language acoustically similar phonemes we can expand the under-resourced acoustic model with the estimated acoustic models from the well-resourced language. Finding a cross-language phoneme mapping table is vital for such an implementation. Since obtaining a table of cross-language similar phonemes is a challenging task, especially when there is no access to a linguist expert, with a knowledge of the under-resourced and well-resourced languages, we also propose an automatic approach, based on the acoustical similarities of phonemes. Such an automatic method is welcome, especially for helping to find the similar phonemes when we are trying to build a new language-dependent speech-synthesis or recognition system for an under-resourced language. To evaluate the automatic cross-language phoneme mapping we also involve the phonetic expert to perform a manual phoneme mapping table. We compare both approaches with the evaluation of synthetic voices.

In a further subsection we describe the steps to obtain the synthetic voice for the under-resourced language in de-

tail. We start with the database description, followed by sections of two approaches to obtain the cross-language phoneme mapping. Later on we also describe the proposed speech-synthesis system for the under-resourced language.

2.1 Database description and preparation

For the purposes of this research we present the Slovenian language as the one with few resources for HLT. Since this is not a real description of the Slovenian speech resources in general, as it is presented at *Multilingual Europe Technology Alliance*[2], we reduced the available VNTV Slovenian speech database. Using only the minimized resources with approximately 2 minutes and 20 seconds of speech, we assessed the Slovenian resources used for purposes in this research as poor. Such a minimization imitates the Slovenian language as the poorly resourced language.

We took the part of the VNTV Slovenian speech database labelled in the SAMPA phonetic representation to include data of only one speaker (speaker 02m). The data were additionally minimized by including just 31 phonetically balanced utterances, which correspond to approximately 2 minutes of acoustic data. Since it is almost impossible to obtain the same phonetic coverage from an already developed Slovenian speech database, as it was approximately calculated in [26], the random selection of balanced utterances was focused in recommendations and simple limitations as follows:

- we picked each recorded utterance as a whole unit from the speech database,
- the random selection was performed as long the requirement for 31 utterances had coverage of all the labelled phonemes,
- the random selection takes into account the utterance phoneme diversity, which consequently results in wider diverseness of the selected annotated context.

The English language is on the other hand, a well-resourced language with publicly available speech resources. In this article we used a part of the well-known CMU ARCTIC database labelled in the ARPAbet phonetic alphabet [27]. We used the training sets of awb, bdl, clb, jmk and rms speakers to estimate the English speaker independent acoustic model.

The basic overview of the minimized Slovenian and English speech resources used in our experiments is presented in Table 1.

[2]http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison

Table 1. Basic statistic of the minimized language resources from the VNTV database and the CMU ARCTIC database

Database attribute	**VNTV subset**	**CMU ARCTIC**
language	Slovenian	English
No. of utterances	31	5653
No. of speakers	1	5
No. of allophones	39	41
No. of labelled phones	1972	190331
Lgth. of 16kHZ a. files	2min 21,2 s	5h 15min 4,8s

2.2 Cross-language phoneme mapping

The relation between two languages is in our case obtained with the mapping of the most similar sounds founded in available speech databases and can be presented with a cross-language phoneme mapping table. Speech databases which are developed for speech processing are often defined in different phonetic representations (phonetic alphabets) in ASCII-encoded characters as SAMPA, X-SAMPA, ARPAbet, etc., which allow computer processing and analysis. On the other hand the human readable IPA phonetic alphabet [28] has been designed as a standardized representation of the sounds of oral language. Its symbols are designed as to be uniform across languages. When the cross-language usage of available speech databases is required, there is often a choice of researcher to remap the phonetic representation with the help of cross phonetic alphabet tables. Commonly, there is not a one-to-one cross-language phoneme match, and therefore special attention is dedicated to find the most similar ones.

We assess the difficulties of establishing phoneme mapping relation with the two different approaches. First we employ the linguist expert with the knowledge of both languages and also with the knowledge of the phonetic presentations available in speech databases, and second we propose an automatic approach to find the best phonetic match between the under-resourced and well resourced speech data. Since the first approach is strongly dependant on the linguist expertise it is desired, as described in Sections 1.1, to propose an automatic solution, which can help or even substitute the difficult task of manual cross-language phoneme mapping.

We present a Slovenian-English phoneme mapping converted to IPA human readable phonetic alphabet in Table 2. The first column presents the available IPA phonemes in the minimized speaker-dependent Slovenian speech database. The second column presents the result of an automatic approach to find the acoustically similar phonemes of cross language and cross-speaker databases. The approach is described in detail in Subsection 2.2.2. The third column presents the subjective expert linguist

Table 2. The Slovenian-English phoneme mapping table comparison in the IPA phonetic for an automatic and manual approach and IPA phonetics

VNTV	CMU automatic mapped	CMU manually mapped	IPA mapping
a	ɔ	ɔ	-
aː	ɑ	ɔ	-
ɒː	ɔ	ɒ	-
b	w	b	b
d	d	d	d
e	ɛ	ɛ	-
eː	ɛ	æ	-
ə	ə	ə	ə
ɛː	ɪ	ə	-
ɜː	ə	ə	-
f	v	f	f
g	u	g	g
i	eɪ	ɪ	i
iː	eɪ	i	-
ɪ	ɛ	ð	ɪ
j	ɛ	j	j
k	k	k	k
l	m	ɬ	-
ʎ	n	ɬ	-
m	ŋ	m	m
n	n	n	n
ŋ	ŋ	ŋ	-
o	oʊ	ɑ	-
oː	oʊ	ɑ	-
p	p	p	p
r	r	r	r
s	s	s	s
ʃ	ʃ	ʃ	ʃ
t	t	t	t
ʦ	s	joined t and s	-
ʧ	ʧ	ʧ	ʧ
u	oʊ	ʊ	u
uː	oʊ	u	-
v	oʊ	v	v
w	oʊ	w	w
x	k	h	h
z	z	z	z
ʒ	ʒ	ʒ	ʒ

decisions about the cross-database phoneme similarities described in 2.2.1. The last column presents the cross-lingual phoneme between all the available phonemes in each database that share the same IPA symbol.

2.2.1 Manual cross-language phoneme mapping

The manual cross-language phoneme mapping between the Slovenian and English languages is based on the phonetic systematization of speech databases. The linguist expert mapped the representations of the SAMPA and ARPAbet phonetic sets. Such a table presents a subjective linguist opinion. Since the IPA phonetic alphabet is a well-established phoneme representation we additionally mapped the resulting matches using the IPA recommendations and other available on-line resources in order to obtain the human-readable phonetic table.

We find that the CMU ARCTIC and Slovenian minimized speech database VNTV share 23 of the same IPA symbols as shown in Table 2. From the second column it is evident that the expert linguist did not match the same IPA symbols, especially when mapping the available vocals in databases. The reported mapping table was obtained from the available phoneme sets based on the SAMPA and ARPAbet phonetic alphabets and by additional listening to the sample words with the available specific phoneme. Reported manually mapped phonemes in Table 2 emphasize the difficult and subjective linguist expert decisions, when trying to obtain a reliable mapping table.

2.2.2 Automatic cross-language phoneme mapping technique

To find the most similar phonemes from different language speech databases and different speakers we propose an approach that was adopted from the field of speaker verification. The approach is based on modelling the acoustic space data with Gaussian mixture models (GMMs) that are derived from the initial universal background model (UBM) [29] using a maximum a posteriori (MAP) adaptation [30]. The UBM serves as the initial model for the MAP adaptation. In our case we tried to build the UBM on all the acoustic data and then the language dependent GMMs for each individual phoneme were estimated by performing a MAP adaptation on language-dependent data.

The UBM is prone to an unbalanced data population, i.e., in the speaker verification task the female and male speech data should be balanced, otherwise the obtained UBM would be biased towards the dominant sub population. In our case we try to overcome the problem of language-dependent unbalanced speech data. One of the possible solutions is to develop only a target language-dependent UBM model. Since we disposed of a small amount of the Slovenian speech data and the normalization of the likelihood scores of the language-dependent GMM also gains an important role. The normalization allows us to find phoneme differences or similarities in

the normalized spaces. With the use of small target-language-dependent UBM and MAP adaptation we obtained phoneme GMMs. The prior UBM model is therefore used in terms of the initial initialization of the training process with the MAP adaptation and later on used for additionally normalizing the cross-language phoneme distances.

The training of the UBM was performed using the EM algorithm [31] to estimate the language-dependent GMM densities. The initialization of the UBM training was performed with the Linde-Buzo-Gray hierarchical method [32]. The training of the UBM is performed in an unsupervised manner. We trained all the available unlabelled data from the under-resourced language based on the acoustic features MFCC [33]. For each utterance we calculated the MFCC vector with the HTK toolkit [34]. We obtained feature vectors with a length of 36 features consisting of 12 MFCC coefficients, 12 delta and 12 delta-delta coefficients per frame. The feature extraction was guided by the following parameters: 32-ms-wide Hamming window with a 10-ms frame shift, a low cut-off frequency of 300 Hz and a high cut-off frequency of 7600 Hz and with a pre-emphasis coefficient of 0.97. Additionally, we found that the crucial step when we deal with different speakers and databases recorded in different conditions is the normalization. With the feature-extraction process we also included the zero mean source of the waveform. Secondly, we performed the cepstral mean normalization based on the whole database set in order to obtain the speaker- and language-normalized data sets.

To obtain the relevant statistics and also to avoid minor mistakes in the automatically segmented CMU ARCTIC database, we discard all the phonemes that had less than 3 features vectors, meaning that if the segmented duration of an allophone was less than 30ms, we discarded it. The MAP was only used on mean vector adaptations. The number of optimal Gaussian mixture components was determined experimentally to be 16. We obtained such a value by comparing the obtained cross-language phoneme mapping tables obtained with the same procedure where only the parameter for the Gaussian mixture components was adjusted from 2 to 2056. The determined parameter of 16 Gaussian mixture components can also be related to the small amount of data (approximately 2 minutes) of he under-resourced language. Figure 2 shows the procedure to obtain the language-dependent phoneme GMMs.

To compute the cross language phoneme similarities these phoneme GMMs were compared by computing a cross log-likelihood ratio (CLR) similarity measure. The proposed CLR similarity measure is given in Equation 1. A similar measure is used in the speaker diarization task to find similarities between the segments of the same speaker

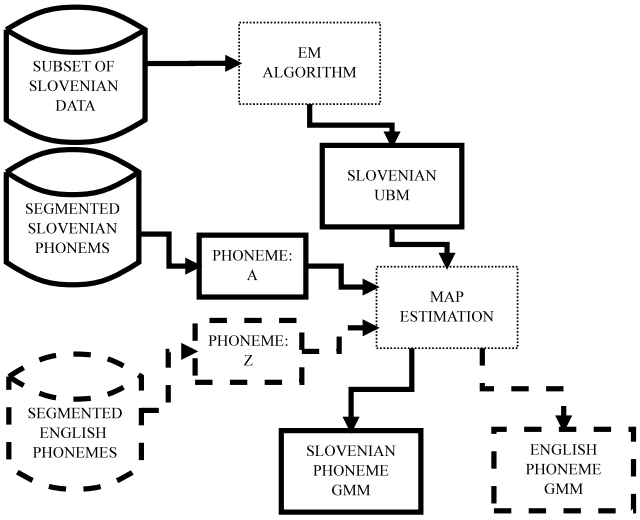

Figure 2. Overview of the proposed process to obtain the language-dependent estimated GMMs

[35].

$$clr(C_i, C_j) = \frac{\log p(x_i|GMM_j)}{\log p(x_i|UBM)} + \frac{\log p(x_j|GMM_i)}{\log p(x_j|UBM)}. \quad (1)$$

C_i and C_j represents the phonemes i and j, while GMM_i and GMM_j are the GMM models estimated from the UBM model with the MAP adaptation on the basis of the x_i feature sample from the phoneme of samples C_i and vice-versa x_j from phoneme C_j. The $clr(C_i, C_j)$ represents the sum of the two log-likelihood ratios. With the first one we can check the affiliation of the cluster C_i to the model GMM_j and with second one the affiliation of the cluster C_j to the model GMM_i. If the CLR value is higher, it is more likely that the clusters C_i and C_j affiliate to similar models of phonemes in different languages. If we perform the calculation of all the phoneme-labelled densities and sort them by the CLR, we obtain the ordered list of the most similar phonemes. The phoneme-labelled density with the highest CLR represents the best match. By exposing its label we can map the phoneme between the Slovenian and English phonetic segmentations. The result of such a process is presented in Table 2 as the mapping between the labelled speech segment from the English CMU ARCTIC database and the minimized Slovenian VNTV databases.

2.3 Design of a cross-language mapped synthetic voice

The speech synthesis is performed with the training, cross-model labelling algorithm, an adaptation of joined model and the synthesis step, as shown in the Fig. 1. First, an estimate of the language-dependent acoustic model is

required. Second, the mapping algorithm of the estimated densities of English language-dependent acoustic model is activated. Third, the joined acoustic model of the Slovenian and English estimated densities is obtained. The joined acoustic model is adapted to the Slovenian data, the purpose being to obtain a speech synthesis system with similar characteristics to the Slovenian speaker.

For all the developed systems we used the HTS speech synthesis toolkit [10], with the same acoustic features and model structure. This research is focused in triphone context-dependent labels for the speech synthesis [36]. The acoustic features consisted of 25 mel-generalised cepstrum coefficients (MGCs) [37], their first-order and second-order deltas, log F0 and their first-order delta, and the energy. The features were extracted using the Hamming window with a 5ms shift. The HMMs were composed of five states with no-skip left-to-right transitions, with one Gaussian mixture for each state. The MSD-HMMs [38], were used for the F_0 modelling. The Mel Log Spectrum Approximation (MLSA) filter [39] was used for the synthesis from the generated speech parameters.

The minimized Slovenian VNTV speech database consists of data from only one speaker. Therefore, the speaker dependent speech-synthesis system was performed. With the well-resourced speech repositories, which is in our case the English language, for the purposes of this research we used 5 different speakers as shown in Table 1. Therefore we performed the speaker adaptive training to obtain the English speaker-independent acoustic model [40].

Based on the phoneme-mapping table the context-dependent labels are transformed from the well-resourced language acoustic model to under-resourced language phoneme representation with the help of the label transform algorithm. Our implementation outputs the acoustical bilingual joined model, with the mapped labels to the under-resourced language. The algorithm consists of relabelling of the estimated densities of the well-resourced language with the conditional joining function, which disables the joining of the already estimated context-dependent labels from the under-resourced context. In our experiments we focused only on mapping the acoustic estimates of the HMM-states. Therefore, we picked the middle phoneme label of the triphone modelled representation to also map the MSD-HMM continuous estimates of the $\log(F_0)$ densities. With such labelling we obtained mapped HMM states with the acoustic estimates borrowed from the well-resourced language and $\log(F_0)$ estimates from the speaker of the poorly resourced language. During the synthesis part the duration estimated model from the under-resourced Slovenian speech data was employed to obtain more accurate, comparable, under-resourced synthetic speech utterances. Table 3 shows the quantity of estimated triphones for the language-dependent acoustic model and the joined bilingual acoustic model obtained from the two different cross-lingual phoneme mapping approaches.

Table 3. Quantities of triphones estimated in different acoustic models; DM - speaker-dependent, SI - speaker independent, MM - exert manual mapping, AM - automatic mapping

Type of acoustic model	**NO. of estimated triphones**
VNTV SD	1113
CMU SI	9975
CMU+VNTV MM	12539
CMU+VNTV AM	10966

After obtaining such a joined model the additional adaptation with the under-resourced language data is performed. The CMLLR adaptation based on regression tree [41] is performed to obtain the speaker characteristic from the under-resourced language. With such a speech synthesise model we are able to synthesis speech in the under-resourced language with the Slovenian speaker characteristics.

3 EVALUATION

Obtaining evaluation results for the under-resourced language often involves a subjective evaluation method. Objective measures, for instance, mel cepstral distortion (MCD) [9] or the root mean square error (RMSE) [42], can provide repeatable experiments, but in real case scenarios where the under-resourced language speech-synthesis system is developed, there is often no available data needed to obtain the statistically reliable objective quantity measure. To test our hypothesis described in Section 1.2, we prepared a special test set of input speech sentences, which represented the most diverse context-dependent labels. Such synthesised utterances in the Slovenian target language consisted of triphones, which were mostly not estimated from the Slovenian speech database, but synthesised with the help of the phonetic tree clustering algorithm.

The test set includes 30 randomly selected sentences from a Slovenian novel. Before the random selection the automatic grapheme-to-phoneme transformation of each sentence from the text of the whole novel was performed. Each grapheme sentence was attributed with the percentage of context-dependent labels that could be estimated from the minimized VNTV database. From subsets of sentences between the affirmation of 20% to 50% only those that had 6-8 words were randomly selected. With such limitations we tried to obtain sentences that have a representative quantity of unseen context-dependent labels, on one

side, and on the other side, give the evaluators the opportunity to concentrate on differences between the sentences that are not too long, and therefore easy to remember. The randomly picked sentences were synthesised with three different speech-synthesis systems. The AB test was performed between 30 sentences and three different speech-synthesis systems. The first and the second were developed using the proposed procedure in Section 2.3. The difference between the first two was in the proposed mapping algorithm. The first speech-synthesis system presents the joined cross-language resources based on the expert phoneme mapping table and the second presents the approach for automatic cross-language phoneme mapping. The last speech-synthesis system presents the baseline system and it is developed only from the minimized Slovenian speech database.

3.1 Evaluation procedure

We implemented the subjective evaluation with the self-developed online web interface presented in Fig. 3. Each participant marked 10 utterances from each system,

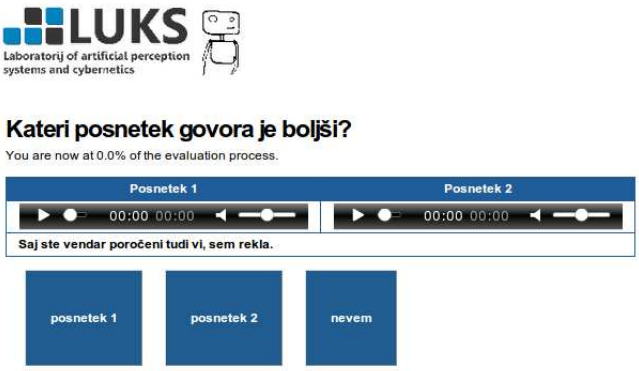

Figure 3. Screenshot from web interface, when evaluator evaluated the synthesised utterances

which were randomly chosen from a pre-specified list of available test utterances. We integrated the validation algorithm, that each synthesised utterance had an equally distributed position, at the time when the evaluator evaluated the utterances. When the evaluator finished the evaluation process each synthesised utterance from each system was positioned five times in position 1 and five times in position 2. With such an implementation we removed the biased decisions of the evaluator opinion towards the last listened utterances. Each evaluator marked a total of the 30 utterances with a simple mouse click on one of the buttons below the players of the synthesised utterances, which were marked as recording 1 and recording 2. The real identity of synthesis system was never revealed to the evaluators. The evaluators had the opportunity to listen to every utterance as many times as they wished. To ensure that the evaluator understands the meaning of the synthesised utterance and also to give the evaluator the option to have an insight into the test text reference, we include the text transcription below the player utterances.

There were three available options to make a decision about the listened to utterances. The first button indicated that the first utterance is better than the second one, the second button indicated the opposite decision. The third opinion indicates that there was no decision about which of the synthesised utterances represented a better speech quality.

The application also consists of the validation if utterances were actually played. For each evaluation step there is also integration of the time measurement of the evaluator"s time spent for registering the opinion. An evaluation of the time measurement enable us to statistically determine the problematic utterances. From a statistical analysis of the durations spent for making a decision we can discuss two assumptions. A long time for the decisions indicates that the evaluator was distracted or that the evaluator did not instantly detect the better synthesised utterance. From such an analysis we did not find any larger deviations.

Even though, the evaluation could be conducted from every where that a computer is available with access to the internet, for purposes of this research the evaluation process was conducted in a silent room with 11 available computers. Eleven students from Faculty of Arts from the University of Ljubljana, who are mainly studying linguistics and had no previous experiences with speech-synthesis systems, were employed for the task of the evaluation. All the evaluators were able to evaluate the speech-synthesis system simultaneously with the use of headphones. Since the evaluation web interface enables randomized access to the testing utterances, no evaluator was presented with same test utterances at the same time.

4 EXPERIMENTAL RESULTS

The evaluation results are presented in Fig. 4, 5 and 6. Each figure presents the evaluator's opinions about the over all quality of the utterances synthesised with three different speech-synthesis systems. As described in Subsection 3.1, the evaluators evaluated the utterances of synthesised speech by comparing two of the same utterances from different speech-synthesis systems (AB subjective evaluation). The evaluator's opinions were collected and divided into three classes: undecided, system A and system B for each system comparison. The collected comparisons were tested for its significance with the binomial proportional test [43]. Since the data is just about the 100 evaluations,

we average the binomial distribution to the normal distribution. The significance level is represented by the significance value of 0.5.

Figure 4 presents the evaluation results from the speech-synthesis system developed with the automatic phoneme table (AM system) and the baseline speech-synthesis system (UR system). From the evaluation results it is evident that our implementation of the automatic cross-language phoneme mapping technique did not improve the quality of the synthesised speech. The baseline system can produce a significantly better synthesised voice with a significance value of $p < 10^{-3}$.

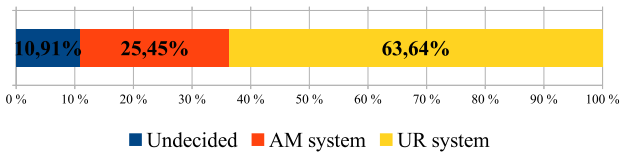

Figure 4. Results obtained from a comparison of the automatic phoneme mapping (AM system) and baseline (UR system) speech-synthesis systems. Significance value $p < 10^{-3}$, n=11, questions=10.

Figure 5 presents the evaluation of the AB subjective test of the speech-synthesis system developed with the proposed expert cross-language phoneme-mapping table (MM system) and the baseline system (UR system). The results suggest, that the speech-synthesis system with joined a Slovenian-English acoustic model produces a better synthesised speech quality. The statistical test of the significance showed that the quality of the produced speech is significantly better than the quality of the baseline system with a p value of 0.038.

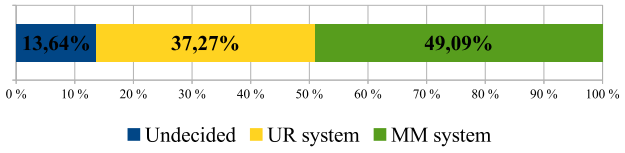

Figure 5. Results obtained from a comparison of the phonetic expert phoneme mapping (MM system) and the baseline (UR system) speech-synthesis systems. Significance value 0.038, n=11, questions=10.

Figure 6 presents the results of the evaluation of produced artificial speech quality obtained from speech-synthesis systems developed with the proposed method for expanding under-resourced HMM acoustic models. The difference between the implementations is only in the phoneme-mapping algorithm. The red bar represents the automatic approach to phoneme mapping (AM system) and the yellow bar represents the cross-language phoneme-mapping based on the expert phoneme mapping (MM system). The results are tested for significance, where we obtained a significance value of 0.015. We can say that the speech-synthesis system developed with the help of the expert cross-language phonetic mapping produces a significantly better quality of artificial speech than the speech-synthesis system developed with a joined acoustic model based on an automatic cross-lingual phoneme-mapping approach.

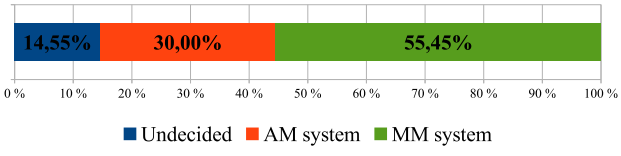

Figure 6. Results obtained from comparison of the automatic phoneme mapping (AM system) and the expert phoneme mapping (MM system) speech-synthesis systems. Significance value 0.015, n=11, questions=10.

5 DISCUSSION

The evaluation results in Fig. 5 confirm our hypothesis, that it is possible to improve the target language acoustic model with the additional speech data from foreign well-resourced language. The improvement is not outstanding, but significant. Our improvements were achieved with use of an expert phoneme-mapping table of phoneme segments between the target under-resourced language and the well-resourced language. When the automatic mapping was applied to the speech-synthesis system the improvement from the subjective evaluation is not detected. From the obtained evaluation results it is evident, that for improving the speech synthesis quality with the proposed implementation of mixing well-resourced English speech data and under-resourced Slovenian speech data, a good and reliable phoneme-mapping table is required. From detailed observations of Table 2 it can be seen that the automatic mapping approach provides some mapping errors. One of the possible causes of these errors can be found from the UBM training. In our implementation of under-resourced language UBM we used the only available acoustic material (2.35 minutes of speech), which was later used for the development of an acoustical model for speech synthesis. The UBM training is performed in an unsupervised manner, meaning that no transcriptions of the processed speech are required. Therefore, the data available for UBM training could be expanded with not

very much additional effort and it is expected that the estimated UBM should be much more biased towards a universal language representation. With such a UBM the estimated language-dependent GMMs with MAP training would probably have wider phoneme distances and consistently more accurate cross language phoneme mapping could be obtained. Wider phoneme distances could also help to determine the threshold for accepting or rejecting the phoneme match. The results obtained with the proposed automatic phoneme mapping technique reports one-to-one cross-language phoneme match based on the highest CLR. If the accurate threshold for the CLR could be determined, we could improve the matching table by rejecting the phoneme matches, that do not provide convincing CLR. Different phone-mapping table would allow us to map only the similar speech segments and also to reject the less similar matches. In that way we could additionally control the estimations of the under-resourced language acoustic models and hopefully obtain better quality of the synthesised speech in the under-resourced language.

The improvements should also be obtained with additional corrections of the expert phoneme mapping. Here, the linguistic expert will need to be more focused on the cross language IPA allophone representation as showed in Table 2. The additional reasons for relatively minor improvements of the acoustic models can also be in the assortment of the target (Slovenian language) and well-resourced language (English language). Such a choice was made due to the fact that the English language is well-resourced and has publicly available speech databases and therefore is one of the most suitable choices to report phoneme mapping strategies for improving the acoustic models of an under-resourced language. On the other hand, such mapping represents a difficult task because of the language group differences (Slavic and German language groups), consequently suggests to the consistent differences in the phoneme and allophones definitions. For example, in Slovene language there is no representation of diphthongs as basic units, and further on the frequently used Slovene phoneme ts in English language does not even exist. We can expect more significant improvements in cases of using similar language group's speech databases, for example Croatian [44] or Serbian, or at least in the cases of phoneme mapping of languages with a similar systematization of basic phoneme units, as shown in the case of bilingual speech-synthesis system for the Slovenian and Croatian languages [11].

6 CONCLUSION

The proposed implementation of a speech-synthesis system for an under-resourced language is one of the possible steps towards low-cost language portability for a synthesis system of the VUI enabled devices. It is based on the HMM-based synthesis system, which is widely used for obtaining polyglot synthesis and is also one of the most suitable techniques for obtaining synthetic speech for under-resourced languages. Our implementation demands a cross-language phoneme-mapping table. With such a table a better acoustic model for synthesised speech in the under-resourced language can be obtained, as shown for the example with the expert phoneme mapping. The proposed automatic approach to find the most similar basic units between the well-resourced and the under-resourced speech databases, evaluated with a subjective evaluation did not improve the baseline speech-system for the under-resourced language. The possible solutions and observations for such a reason are discussed in Section 5. From the obtained results it is evident that our proposed technique is strongly dependent on well defined cross-lingual phoneme mapping. Since the speech synthesis with the expert phoneme mapping did provide a significantly better speech quality, we found that the approach for cross-lingual mapping of well-resourced and under-resourced language acoustic models is a suitable method to obtain better acoustic coverage for under-resourced language. On the other hand, there is still space for improvements to the proposed automatic approach for the cross-lingual phoneme-mapping technique, which can provide the mapping of under-resourced and well-resourced speech-segment similarities. The automatic approach is a convenient method for performing low-cost language portability with a limited amount of data. Therefore, additional experiments and investigations should be performed, not only with the proposed technique, but also with other possible automatic approaches, which we will try to include in our future work. Also, the language differences gain an important role therefore we will investigate the proposed technique with closely related languages.

ACKNOWLEDGEMENTS

The work presented in this paper was supported by junior research grand founded by Slovenian Research Agency (ARRS) with grant number 1000-09-310132 and in parts by the national research program P2-0250(C) Metrology and Biometric Systems, the European Union's Seventh Framework Programme (FP7-SEC-2011.20.6) under grant agreement number 285582 (RESPECT). The authors additionally appreciate the support of COST Actions IC1106 and IC1206.

References

[1] M. H. Cohen, *Voice user interface design*. Addison-Wesley Professional, 2004.

[2] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A

survey ," *Speech Communication*, vol. 56, no. 0, pp. 85 – 100, 2014.

[3] H. Lin, J.-t. Huang, F. Beaufays, B. Strope, and Y.-h. Sung, "Recognition of multilingual speech in mobile applications," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012*, pp. 4881–4884, IEEE, 2012.

[4] V.-B. Le and L. Besacier, "Automatic Speech Recognition for Under-Resourced Languages: Application to Vietnamese Language," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1471–1482, 2009.

[5] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[6] J. Žibert and F. Mihelič, "Slovenian weather forecast speech database," in *Proc, SoftCOM*, vol. 1, pp. 199–206, SoftCOM, 10 2000.

[7] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1, pp. 373–376 vol. 1, May 1996.

[8] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis ," *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.

[9] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality on new languages calibrated with mel-cepstral distorion," in *in SLTU 2008, Hanoi, Viet Nam*, 2008.

[10] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. of Sixth ISCA Workshop on Speech Synthesis*, pp. 294–299, 2007.

[11] T. Justin, M. Pobar, I. Ipšić, F. Mihelič, and J. Žibert, "A bilingual HMM-based speech synthesis system for closely related languages," in *Text, Speech and Dialogue*, pp. 543–550, Springer Berlin Heidelberg, 2012.

[12] J. Dijkstra, L. C. Pols, and R. J. v. Son, "Frisian TTS, an example of bootstrapping TTS for minority languages," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[13] N. T. Vu, F. Kraus, and T. Schultz, "Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 5000–5003, May 2011.

[14] T. Schultz and A. Waibel, "Multilingual and Crosslingual Speech Recognition," in *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, pp. 259–262, 1998.

[15] K. C. Sim and H. Li, "Robust phone set mapping using decision tree clustering for cross-lingual phone recognition," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4309–4312, March 2008.

[16] C. Traber, K. Huber, K. Nedir, B. Pfister, E. Keller, and B. Zellner, "From multilingual to polyglot speech synthesis," in *Proc. of the Eurospeech*, vol. 99, pp. 835–838, 1999.

[17] J. Latorre, K. Iwano, and S. Furui, "New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer," *Speech Commun.*, vol. 48, no. 10, pp. 1227 – 1242, 2006.

[18] M. Pobar, T. Justin, J. Žibert, F. Mihelič, and I. Ipšić, "A Comparison of Two Approaches to Bilingual HMM-Based Speech Synthesis," in *Text, Speech, and Dialogue*, pp. 44–51, Springer Berlin Heidelberg, 2013.

[19] T. Schultz, N. Vu, and T. Schlippe, "GlobalPhone: A multilingual text amp; speech database in 20 languages," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 8126–8130, May 2013.

[20] Y. Qian, H. Liang, and F. Soong, "A Cross-Language State Sharing and Mapping Approach to Bilingual (Mandarin-English) TTS," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, pp. 1231–1239, Aug 2009.

[21] X. Cui, J. Xue, X. Chen, P. Olsen, P. Dognin, U. V. Chaudhari, J. Hershey, and B. Zhou, "Hidden Markov Acoustic Modeling With Bootstrap and Restructuring for Low-Resourced Languages," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, pp. 2252–2264, Oct 2012.

[22] Y. Qian, J. Xu, and F. Soong, "A frame mapping based HMM approach to cross-lingual voice transformation," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 5120–5123, May 2011.

[23] H. Cao, T. Lee, and P. Ching, "Cross-lingual speaker adaptation via Gaussian component mapping.," in *INTERSPEECH*, pp. 869–872, 2010.

[24] S.-J. Kim, J.-J. Kim, and M. Hahn, "HMM-based Korean speech synthesis system for hand-held devices," *IEEE Trans. Consumer Electronics*, vol. 52, pp. 1384–1390, Nov 2006.

[25] J. Žganec Gros and M. Žganec, "An efficient unit-selection method for embedded concatenative speech synthesis," *Informacije MIDEM - Journal of Microelectronics, Electronic Components and Materials*, vol. 37, no. 3, pp. 158–164, 2007.

[26] F. Mihelič, J. Gros, J. Dobrišek, S.and Žibert, and N. Pavešić, "Spoken Language Resources at LUKS of the University of Ljubljanai," *International Journal of Speech Technology*, vol. 6, no. 3, pp. 221–232, 2003.

[27] D. H. Klatt, "Review of the ARPA speech understanding project," *The Journal of the Acoustical Society of America*, vol. 62, no. 6, pp. 1345–1366, 1977.

[28] I. P. Association and C. A. I. Corporate, *Handbook of the International Phonetic Association : A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, June 1999.

[29] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using Adapted Gaussian mixture models," in *Digital Signal Processing*, p. 2000, 2000.

[30] J. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 291–298, Apr 1994.

[31] A. P. Dempster, N. M. Laird, D. B. Rubin, *et al.*, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[32] Y. Linde, A. Buzo, and R. Gray, "An Algorithm for Vector Quantizer Design," *Communications, IEEE Transactions on*, vol. 28, pp. 84–95, Jan 1980.

[33] E. Standard, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms," tech. rep., ETSI, 2003.

[34] S. Young and S. Young, "The HTK Hidden Markov Model Toolkit: Design and Philosophy," *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.

[35] J. luc Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, vol. 37, pp. 89–108, 2002.

[36] M.-Y. Hwang and X. Huang, "Subphonetic modeling with Markov states-Senone," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1, pp. 33–36 vol.1, Mar 1992.

[37] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-Generalized Cepstral Analysis," in *Proc. ICSLP-94*, pp. 1043–1046, 1994.

[38] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 1, pp. 229–232 vol.1, Mar 1999.

[39] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.

[40] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, pp. 1208–1230, Aug 2009.

[41] M. J. Gales, *The generation and use of regression class trees for MLLR adaptation*. University of Cambridge, Department of Engineering, 1996.

[42] A. Vasilijević and D. Petrinović, "Perceptual Significance of Cepstral Distortion Measures in Digital Speech Processing," *AUTOMATIKA: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije*, vol. 52, no. 2, pp. 132–146, 2011.

[43] R. B. D'agostino, W. Chase, and A. Belanger, "The Appropriateness of Some Common Procedures for Testing the Equality of Two Independent Binomial Populations," *The American Statistician*, vol. 42, no. 3, pp. 198–202, 1988.

[44] S. Martinčić-Ipšić, M. Pobar, and I. Ipšić, "Croatian large vocabulary automatic speech recognition," *AUTOMATIKA: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije*, vol. 52, no. 2, pp. 147–157, 2011.

Tadej Justin Tadej Justin was born in 1983 in Ljubljana, Slovenia. In 2009 he obtained his B.Sc degree in electrical engineering from the University of Ljubljana, Faculty of Electrical Engineering. He currently works as Researcher at University of Ljubljana, Faculty of Electrical Engineering. His research interests include statistical modeling, emotional speech synthesis, emotional speech recognition and cross-language related speech technologies with a special focus on the Slovenian language.

France Mihelič France Mihelič studied at the Faculty of Natural Sciences, Faculty of Economics and Faculty of Electrical Engineering all at the University of Ljubljana. There he received the B.Sc. degree in Technical Mathematics, the M.Sc. degree in Operational Research and the Ph.D. degree in Electrotechnical Sciences in 1976,1979 and 1991, respectively. Since 1978 he has been a staff member at the Faculty of Electrical and Computer Engineering in Ljubljana, where he is Full Professor, and the Head of the Laboratory for Artificial Perception, Systems and Cybernetics. His research interests include Pattern Recognition, Speech Recognition and Understanding, Speech Synthesis and Signal Processing.

Janez Žibert Janez Žibert received his B.Sc. degree in mathematics in 1998 and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Ljubljana in 2001 and 2006, respectively. He is currently working as an Associate Professor at the Faculty of Health Sciences at University of Ljubljana and as a Research Fellow at the 'Andrej Marušič' Institute at University of Primorska. His research interests include statistical modeling, pattern recognition, machine learning in general with focus on audio-signal and image data processing.

AUTHORS' ADDRESSES

Tadej Justin, B.Sc.
Prof. France Mihelič, Ph.D.
Laboratory of Artificial Perception, Systems and Cybernetics (LUKS),
Faculty of Electrical Engineering,
University of Ljubljana,
Tržaška 25, SI-1000 Ljubljana, Slovenia
email: tadej.justin@fe.uni-lj.si, france.mihelic@fe.uni-lj.si

Assoc. Prof. Janez Žibert, Ph.D.
Faculty of Health Sciences,
University of Ljubljana,
Zdravstvena pot 5, SI-1000 Ljubljana, Slovenia
email: janez.zibert@zf.uni-lj.si

Received: 2014-10-28
Accepted: 2015-05-04