ISSN 0005-1144
ATKAFF 51(1), 79–88(2010)

Branimir Dropuljić, Davor Petrinović

Development of Acoustic Model for Croatian Language Using HTK

UDK 004.934'1:811.163.42
IFAC 2.8.3; 5.0

Original scientific paper

Paper presents development of the acoustic model for Croatian language for automatic speech recognition (ASR). Continuous speech recognition is performed by means of the Hidden Markov Models (HMM) implemented in the HMM Toolkit (HTK). In order to adjust the HTK to the native language a novel algorithm for Croatian language transcription (CLT) has been developed. It is based on phonetic assimilation rules that are applied within uttered words. Phonetic questions for state tying of different triphone models have also been developed. The automated system for training and evaluation of acoustic models has been developed and integrated with the new graphical user interface (GUI). Targeted applications of this ASR system are stress inoculation training (SIT) and virtual reality exposure therapy (VRET). Adaptability of the model to a closed set of speakers is important for such applications and this paper investigates the applicability of the HTK tool for typical scenarios. Robustness of the tool to a new language was tested in matched conditions by a parallel training of an English model that was used as a baseline. Ten native Croatian speakers participated in experiments. Encouraging results were achieved and reported with the developed model for Croatian language.

Key words: Acoustic model, Automatic speech recognition, Croatian language, Hidden Markov models, Phonetic assimilation, Phonetic transcription algorithm, Recognition accuracy

Razvoj akustičkog modela hrvatskog jezika pomoću alata HTK. Rad opisuje razvoj akustičkog modela hrvatskog jezika za potrebe sustava za automatsko prepoznavanje govora. Prepoznavanje prirodnog spojenog izgovora ostvaruje se korištenjem skrivenih Markovljevih modela (HMM) u okviru alata HTK. U svrhu prilagodbe ovog alata na hrvatski jezik razvijen je novi algoritam za automatsku fonetsku transkripciju hrvatskih riječi. Zasniva se na načelu fonetske asimilacije unutar izgovorenih riječi. Razvijen je i skup fonetskih pitanja koji se koristi za klasifikaciju prilikom udruživanja trifonskih modela sličnih glasova. Razvijena je automatizirana aplikacija za gradnju i evaluaciju akustičkih modela, integrirana s novo razvijenim grafičkim sučeljem. Primjene ovog sustava za prepoznavanje su trening s doziranim izlaganjem stresu (SIT) i terapija izlaganjem primjenom virtualne stvarnosti (VRET). Prilagodljivost akustičkog modela na zatvoren skup govornika vrlo je važna za takve primjene, pa se u radu istražuje primjenjivost alata HTK u tipičnim scenarijima. Robusnost alata na promjenu jezika istražuje se uparenim treniranjem i evaluacijom ekvivalentnog modela engleskog jezika u jednakim uvjetima. U eksperimentima je sudjelovalo deset izvornih hrvatskih govornika. Ostvareni rezultati za hrvatski jezik prikazani u radu pokazuju zadovoljavajuća svojstva razvijenog akustičkog modela hrvatskog jezika.

Ključne riječi: akustički model, automatsko prepoznavanje govora, hrvatski jezik, skriveni Markovljevi modeli, algoritam za fonetsku transkripciju, fonetska asimilacija, točnost prepoznavanja

1 INTRODUCTION

This article describes an early development of automatic speech recognition (ASR) system for subjects' responses in Croatian language, based on verbal information. As in the conventional ASR the idea lies in using the recorded speech utterance signal to extract the cepstral coefficients linked to the characteristic of the vocal tract and using them as output observations that are modeled by Hidden Markov Models (HMM) [1–3].

Some studies were done on exploring ASR systems for Croatian language. News and weather forecasts spoken on the radio and also weather reports over the telephone were collected in the multi-speaker Croatian speech corpus VEPRAD, which was used for building acoustic system with promising results [4–6]. Much more experiments have been done for Slovenian language. A large multi-speaker Slovene speech database GOPOLIS, derived from real situation dialogs concerning airline timetable information services was used for estimating model parameters [7]. Adaptation of the Slovenian language model for multilin-

gual speech recognition was also made [8].

In spite of significant similarities between the two languages, certain differences exist in phonemes and the way that they are used in constructing words. Therefore, it is not optimal to use the existing Slovenian models. Croatian database VEPRAD would be a good starting point, but we wanted to build and test a new speech corpus that can be applied in creating acoustic and language models associated with expression of emotions and emotional speech. First, it was necessary to build up a new acoustic model for the Croatian language. The HTK tool (version 3.3.) was chosen for this purpose due to its popularity and open availability of sophisticated training tools to build up a new language model from the start [9]. In order to perform a high quality customization of this tool it was necessary to design and implement linguistic and phonetic rules of the Croatian language by creating a dictionary of word transcripts and phonetic questions [10].

The targeted applications of the developed ASR system are stress inoculation training (SIT) and virtual reality exposure therapy (VRET). During such training or therapy, a subject is exposed to different levels of stress situations and system must estimate the emotional state of the subject, or his/her level of stress. Response of the subject is measured using different modalities, like psychophysiological responses (heart rate, blood pressure, skin conductance and temperature), but also through verbal and nonverbal responses. The analysis of vocal features can give important information about the emotional state of a person [11]. Emotions can be expressed non-verbally, through prosodic structure of an utterance, but also verbally, by directly expressing thoughts and feelings. When it comes to emotions, a verbal or a direct form of utterance is often suppressed or hesitated to respect the limits of social tolerability. Provided that a person utters what he or she feels without hesitation in a controlled therapy setting, a very valuable piece of information is obtained that can be used for estimating the emotional state of a person through semantic analysis (e.g. keywords and n-grams) [12,13]. The emotional state estimator is part of an ongoing research project: *Adaptive control of virtual reality scenarios in therapy of posttraumatic stress disorder (PTSD)*, at University of Zagreb [14,15].

In order to obtain the best recognition results, acoustic models must be designed from and optimized for responses of a closed set of subjects (usually one or a few) that are undergoing the targeted therapy. Therefore, our goal was to design an automated ASR training mechanism that enables automatic building of speaker dependent acoustic models performed by trainers or therapists that are not experts in the ASR. Such training must also be robust to a limited data base size acquired from initial sessions. Experiments presented in this paper are intended to verify the applicability of this concept in the intended scenarios.

A paired model for the English language was also designed from scratch to serve as a baseline for comparison [16]. Training and testing of these two systems was performed in matched conditions in order to verify the suitability of the HTK tool for a new language.

2 SETTING UP ACOUSTIC MODELS

As it usually done for continuous speech recognition with unlimited vocabulary, a large HMM net is used. Hence, sub-word units must be used to model the speech. In our setup we have trained triphone, biphone and monophone models. Since the same phonemes in different words are uttered differently depending on their position in the word and their neighboring phonemes, the ASR system based on the largest unit offers the greatest modeling accuracy [9]. For triphone transcriptions, words and sentences are transcribed as a model chain where each model has one central monophone. In further presentation, the left and the right context are separated with minus and plus symbols respectively. The following are examples of a sentence transcription *Ivo is in school* (Cro. *Ivo je u školi*). For transcription using the word-internal technique the sentence becomes: 'sil i+v i-v+o v-o sp j+e j-e sp u sp š+k š-k+o k-o+l o-l+i l-i sil' [9]. Note that this technique utilizes all three types of sub-word units. Alternatively, if the cross-word transcription technique is used, the sentence is transcribed as: 'sil sil-i+v i-v+o v-o+j sp o-j+e j-e+u sp e-u+š sp u-š+k š-k+o k-o+l o-l+i l-i+sil sil' [9]. It can be noticed that for the second technique, transcription uses only triphone models besides the 'sil' and 'sp' models. The 'sil' model represents a longer pause at the beginning and at the end of each sentence, whilst a short pause is represented using a 'sp' model. Described triphone transcriptions represent the final form of the reference that is used to train the system for continuous speech recognition with unlimited vocabulary. Large number of states in different models are tied together to enhance the quality and robustness of the HMM [9].

System training is performed through multiple iterations where model parameters of the system are re-estimated in each iteration using Baum-Welch algorithm [9], [17]. At the initial stage, 32 monophone models are initialized for Croatian language, by means of the so called flat start method; where each of these models represents one phoneme of the Croatian language. Furthermore, each of them is modeled as a state machine that consists of 5 different states oriented from left to right without state skipping, as shown in Fig. 1. The first and the last states represent non-emitting entry and exit states, whilst the central three represent the emitting states of each phoneme. Emit-

ting states have a loop-connection that enables them to remain unchanged as long as it is needed. Silence and pause models have the same structure but with only one emitting state. In further training stages, this monophone model is used to train the biphone and triphone models that are used for described transcription. Structure of these models is the same as for the monophone case (Fig. 1). Thus, initialization of the new models is obtained by simply copying parameters of their central monophone model and state tying using phonetic questions.

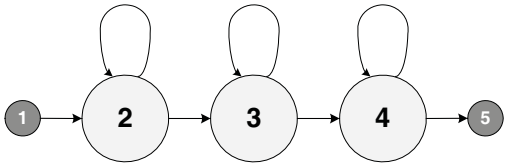

Fig. 1. Phoneme HMM with entry, exit and 3 emitting states

Each emitting state is represented by one Gaussian model with the mean vector and the corresponding covariance matrix. Output observation vectors for each state are 39-dimensional vectors computed from 13 mel-frequency cepstral coefficients (MFCC) [18], 13 delta and 13 acceleration coefficients that are extracted from the recorded speech utterance.

3 ADJUSTING ACOUSTIC MODEL FOR CROATIAN LANGUAGE

In order to develop an acoustic model for Croatian language, it was necessary to define forms of word transcription and different options of phoneme tying.

Standard Croatian language is dialectally closest to the Western Štokavian dialect. Two common dialects that are also used in Croatia are Čakavian and Kajkavian. In our experiment, we have used only the Standard version, with the Ijekavian reflex of the common Slavic yat vowel.

In this article, phonemes are denoted according to the International Phonetic Alphabet (IPA) system using the IPA-SAM phonetic font enclosed in slashes [19].

3.1 Pronouncing Dictionary

An electronic dictionary of word transcripts containing phonetic pronunciations of the entire vocabulary used in the ASR is a necessary prerequisite for working with the HTK. Croatian is mostly a phonemic language with almost one-to-one relation between orthography and pronunciation. With only a few exceptions, transcription is possible by means of a finite set of rules [20], some of which are already used in [5] and [6]. Thus, the creation of a dictionary of Croatian pronunciation can be easily automated; unlike for the English language that requires manual transcription due to complicated pronunciations rules with numerous exceptions. There have been attempts to perform similar automation for the English language as well, but the algorithm is of much higher complexity [21]. The dictionary used in [5] and [6] also includes the stress for each word. Accented and non-accented vowels are differentiated using a colon mark for stressed vowels. This information is not necessary for speech recognition using HTK, but is useful for natural speech synthesis that is also presented in [5] and [6].

Croatian language has 32 phonemes, 30 of which are letters of the alphabet (Table 1.) [22]. There are two sounds that are also considered phonemes. Syllabic form of *r*, /r̩/, is most often found between two consonants having the role of a vowel and the diphthong /ie/ as a substitute to the sequence *ije* in a word, that represents Ijekavian reflex of the common Slavic yat [20], [23]. The phone set used in [5] is identical to the one presented in Table 1. except for the diphthong /ie/ that was not included as a separate phoneme.

Table 1. Croatian phonemes and graphemes

Vowels		**Carriers**		**Consonants**			
				Sonorous		***Non-sonorous***	
P	***G***	***P***	***G***	***P***	***G***	***P***	***G***
/a/	*a*	/j/	*j*	/b/	*b*	/p/	*p*
/ɛ/	*e*	/l/	*l*	/d/	*d*	/t/	*t*
/i/	*i*	/ʎ/	*lj*	/g/	*g*	/k/	*k*
/ɔ/	*o*	/m/	*m*	/dʒ/	*dž*	/tʃ/	*č*
/u/	*u*	/n/	*n*	/dʑ/	*đ*	/tɕ/	*ć*
/ie/		/ɲ/	*nj*	-	-	/ts/	*c*
/r̩/		/r/	*r*	/ʋ/	*v*	/f/	*f*
		/ʋ/	*v*	/z/	*z*	/s/	*s*
				/ʒ/	*ž*	/ʃ/	*š*
				-	-	/x/	*h*

Beside these phonemes, Croatian language has allophones (conditioned realization of the same phoneme). These are mostly the so called complementary allophones generated by phonetic environment, but also free variants due to differences in pronunciation of the same phoneme in the same context by different speakers with different pronunciation habits and styles. According to [23], several allophones (e.g. [ŋ], [ʒ], [ɣ], [ś], [ź], [F]) should be easily distinguishable from their base phonemes by trained listeners. For other allophones, differences in pronunciation are considered to be more subtle. Nevertheless, since the natural transitions within a word are covered with concatenated triphone models, it is not necessary to include allophones as separate models, i.e. complementary allophones are directly modeled using individually trained triphone models.

The number of complementary allophones and thus the number of required triphones is primarily determined by

the number of phonemes of a certain language. For example, the English language has 18 to 23 vowels, and 24 or 25 consonants (according to different sources and types of English language) [24]. For Croatian language with 7 vowels and 25 consonants and carriers, the total number of allophones is much smaller and thus the corresponding acoustical model is more compact and easier to train, what will be demonstrated in our experiments.

Orthography of the Croatian language has changed through the recent history [20]. Nevertheless, the pronunciation and phonetic rules in the Croatian language tend to keep the naturalness and fluency of the speech. The proposed Croatian Language Transcription algorithm (CLT) originates from the idea of preventing possible differences between textual forms of training examples and their pronunciation that can occur due to changes or slight variations of the applied orthography. CLT tries to observe the logic of natural pronunciation by applying phonological and phonetic assimilations irrespectively of their inclusion in the actual orthography. The algorithm is fully described through 6 rules that are given in Table 2.

These rules represent either a positive case (usage/U), or a negative case (exception/E), or even both, as it is shown in the third column.

Table also shows that certain assimilation examples can follow up to two or more different rules. It is therefore recommended to adhere to the ordering in applying these rules in the sequence as they are listed in Table 2. The most frequent combinations are rule 2 followed by 4, or 2 followed by 5. For composite application of rules, the last column of the table gives the required sequence, where * denotes the current rule while numbers denote preceding or succeeding rules.

CLT in the current form can be applied only to the words of Croatian origin or to assimilated foreign words that adhere to phonemic transcription (e.g. *vikend*, *softver*). It is not intended for foreign words in the original orthography, numerical data, dates, acronyms and similar, i.e. it must be preceded with a text normalization algorithm. Furthermore the transcription of compound words (words that are commonly connected in natural pronunciation) is excluded. This is not because the CLT gains on complexity – as the same rules are applied – but only because such approach would generate new words that require adaptation of the dictionary and corresponding word network. In example *I'm passing through a yellow light* (Cro. *Prolazim kroz žuto svjetlo*), the cross-word application of CLT rules would transcribe it as /prˌɔlazim krˌɔʒutɔ svjɛtlɔ/ and thus generate a new word. First solution is to add such newly created word *krožuto* and similar words that do not adhere to Croatian orthography in the dictionary together with the original words *kroz* and *žuto*. As an alternative, recognition and network modeling can be reduced to a sub-word level, while the actual word recognition can be performed from the sequence of recognized sub-word units.

3.2 Phonetic Questions

An important feature of the HTK toolkit is its ability to tie the states belonging to different models into leaf-nodes of a binary tree [9]. This gives a more compact model, but also improves the robustness since several similar triphones are used to train parameters of shared states, which are represented with leaf-nodes. This enables successful training of acoustic models even with limited training material, i.e. for building of triphone models that are not represented in the training database. Branching is based on specific questions linked to the left and right context of the central monophone in a triphone. The questions are formed according to the characteristics and sonority of the pronunciation in the Croatian language and are linked to the manner and place of articulation, sonority of sounds etc.

Since phonetic questions are numerous (cca. 3 pages), they are not presented in this paper. However the list can be found in the Appendix 6 of [10].

4 TRAINING AUTOMATION AND GUI INTERFACE FOR CROATIAN ACOUSTIC MODEL BUILDING

HTK toolkit is based on a large number of executable files, which are invoked through a command line interface. Each of these executables accepts a large number of input parameters, that must be configured properly [9]. Since the tool is highly configurable, significant effort is needed to understand all the possible options offered by these tools. Training and testing can be automated by using scripting languages. In order to simplify this to the maximum extent, we have developed a Graphical User Interface (GUI) in Matlab for training, testing and results' analysis, as shown in Fig. 2, [10]. The developed program achieves complete automatization of all necessary steps that are required in training and evaluation of the ASR system. This is very important for the end-users that are not technical experts in this field. Still it enables sufficient customization to a particular task through the GUI interface (e.g. selection of training and testing databases, HMM models, languages and the most significant ASR parameters). The equivalent GUI environment was developed for English language as well [16].

5 TESTING AND COMPARISON OF ACOUSTIC MODELS

Training and testing of Croatian and English models were done in matched conditions. Ten native Croatian

Table 2. Croatian transcription rules

Rule Num.	Rule	Usage / Exceptions		Example		
				Word	***Transcription***	
1	Omitting of consonants *t* and *d* from *st*, *zd* and *žd* if followed by a consonant	U	foreign words	*rostfraj*	/rɔsfraj/	
		U	feminine nouns deriving from masculine nouns ending in *ist*	*feministkinja*	/fɛminiskiɲa/	
		U	on a crossing of the compound words	*postdiplomski*	/pɔsdiplɔmski/	*2
		E	following consonant is *v*, *j* or *r*	*bratstvo*	/bratstʋɔ/	*2
		E	sequence is from the first syllable (not implemented)	*istkati*	/iskati/ (should be /istkati/)	
2	Modification of the sequence of consonants into sonorous/non-sonorous pairs, depending on the sonority of the last consonant (Table 1.)	U	sequence of two consonants	*postdiplomski*	/pɔzdiplɔmski/	1*
		U	sequence of three or more consonants	*predstava*	/prɛtstaʋa/	*5
		U	modification of one sonorous / non-sonorous pair	*subpolaran*	/suppɔlaran/	*4
		E	consonant *v* is the last in a sequence – in this case it assumes a role of carrier and the rule is not abided by (Table 1.)	*bratstvo*	/bratstʋɔ/	1*5
3	Pacing of the phoneme /j/ between a pair of two vowels where at least one of them is either *i* or *e*	U		*mie*	/mijɛ/	*5
4	Conversion of more consecutive phonemes in one phoneme	U		*subpolaran*	/supɔlaran/	2*
5	Combining two or more graphemes in a sequence of one or more phonemes in particular situations	U	*tc* → /ts/ (*c*)	*bitci*	/bitsi/	
		U	*ts* → /ts/ (*c*)	*predstava*	/prɛtstaʋa/	2*
				bratstvo	/bratstʋɔ/	2*
		U	*tč* → /tʃ/ (*č*)	*mlatče*	/mlatʃɛ/	
		U	*ć* → /tɕ/ (*ć*)	*odćarlijati*	/ɔtɕarlijati/	2*
		U	*š* → /tʃ/ (*č*)	*predškolski*	/prɛtʃkɔlski/	2*
		U	*dz* → /ts/ (*c*) (*tsb* → /dzb/ → /tsb/)			2*
		U	*dž* → /dʒ/ (*dž*)	*sladoleddžija*	/sladɔlɛdʒija/	
		U	*đ* → /dʑ/ (*đ*)	*poddakon*	/pɔdʑakɔn/	
		U	*sš* → /ʃ/ (*š*)	*uzšetati*	/uʃɛtati/	2*
		U	*zž* → /ʒ/ (*ž*)	*razžvakati*	/raʒʋakati/	
		U	*sč* → /ʃtʃ/ (*šč*)	*rasčlaniti*	/raʃtʃlaniti/	
		U	*sć* → /ʃtɕ/ (*šć*)			
		U	*zdž* → /ʒdʒ/ (*ždž*)			
		U	*zđ* → /ʒdʑ/ (*žđ*)	*razđakoniti*	/raʒdʑakɔniti/	
		U	*np* → /mp/ (*mp*)	*jedanput*	/jɛdamput/	
		U	*nb* → /mb/ (*mb*)	*stanben*	/stambɛn/	
		U	*nm* → /m/ (*m*)			
		U	*cč* → /tʃ/ (*č*)			
		U	*cć* → /tɕ/ (*ć*)			
		U	*ije* → /ie/ (diphthong)	*mie*	/mie/	3*
6	Applying phoneme /ṛ/ in particular situations	U	consonant *r* is at the beginning of a word and after it there is another consonant	*rzati*	/ṛzati/	
		U	*r* is at the end of the word and in front of it there is a consonant	*žanr*	/ʒanṛ/	
		U	*r* is between two consonants	*prst*	/pṛst/	
		E	*r* is between two consonants and consonant *j*, *r*, *l*, *lj*, *n*, *nj*, *ć*, *dž* or *đ* is in front of *r*			
		E	before *r* there is consonant and after is a vowel *o*	*istro*	/istṛɔ/	

speakers (5 men and 5 women), aged 20 to 25, took part in experiments. All the speakers were fluent in English such that the same 10 speakers were used for recording of the English materials as well.

Random texts in Croatian and English language were taken from the Internet for training and testing of these models and 500 sentences were formed and recorded for each language. The sentence length was limited to a maximum of 100 characters. All words and signs not predicted as references were removed. Bilingual speech corpus was formed as shown in Table 3. by reading and recording these sentences. The first set was used to train and test a speaker dependent system customized to only one speaker (BD – for Croatian and NM – for English). In this case,

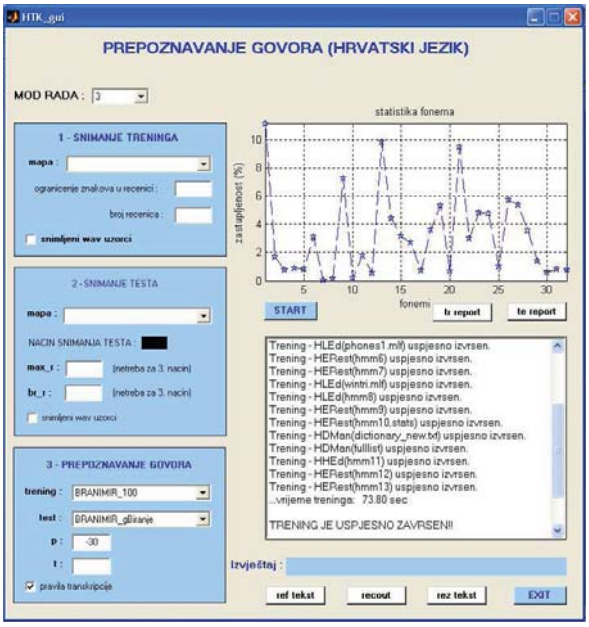

Fig. 2. GUI interface in Matlab for HTK for Croatian language

a single speaker uttered all 500 sentences. The second set was used for speaker independent models. For this set, all ten speakers uttered 50 sentences each in both languages. Although the minimum required number of speakers for building a true speaker independent model is around 50, with the proper balance between gender, age and dialects [25], this limited data set was used in the experiment that verifies the worst case ASR performance for a closed set of speakers. The third set for both languages was uttered by the same speakers as in the first set but for this set simply structured phone dialing sentences were used intended for testing of the grammar influence.

The accuracy of correct transcription of all recorded datasets was verified manually prior to training.

Adjusting an ASR system for a new subject in psychotherapy usually means dealing with the scarce data for building a new speaker dependent system from initial sessions. Hence, a logical path is to iteratively customize the existing speaker independent system to a new speaker until the data sets are large enough to create individually trained ASR system for this particular speaker. Also, the limited number of speakers in the initial phase of building a new ASR system for psychotherapy can be a problem. Therefore, our intention was to verify the accuracy of the acoustic models in such resource-constrained conditions with several experiments that are explained next.

Table 3. Bilingual speech corpus

Croatian Language			
Set Name	C_1_500	C_10_500	C_1_50_pd
Speakers (num.)	1 (BD)	10	1 (BD)
Sentences (num.)	500	500	50
Set Type	random text	random text	phone dialing
Duration (hh:mm:ss)	00:43:05	00:37:21	00:02:54
Dictionary (num. of words)	2403	2403	23
Recording Param.	16 KHz, 16 bit mono	16 KHz, 16 bit mono	16 KHz, 16 bit mono
English Language			
Set Name	E_1_500	E_10_500	E_1_50_pd
Speakers (num.)	1 (NM)	10	1 (NM)
Sentences (num.)	500	500	50
Set Type	random text	random text	phone dialing
Duration (hh:mm:ss)	00:36:04	00:30:52	00:02:53
Dictionary (num. of words)	1954	1954	26
Recording Param.	16 KHz, 16 bit mono	16 KHz, 16 bit mono	16 KHz, 16 bit mono

5.1 Speaker Dependent Models

As a baseline model for comparison and testing, a speaker dependent acoustic model was created for both languages in the first experiment. The first set from the speech corpus was used for creating this particular model (C_1_500 and E_1_500). Sentences from these two databases were separated into two separate sets for training and for testing with the ratio of 90:10. Word-internal method and the proposed CLT algorithm were applied for the Croatian transcription. For evaluation of the trained model, no language model or grammar were used, thus the performance of the ASR was solely determined by the accuracy of the trained acoustic models. This actually corresponds to an unigram language model with equal probability of all words (i.e. the so called flat model) denoted with *Guf* abbreviation in Table 4. Hence, word networks with approximately 2000 words were created from 500 sentences for both languages.

The results of all performed experiments are summarized in Table 4. in the form of word correctness, (N-D-S)/N * 100%, and word accuracy, (N-D-S-I)/N * 100%, where N represents the number of words in the testing set, D is the number of deleted words, whereas S stands for substituted and I for the number of inserted words. The obtained word correctness was 90% for the Croatian language and 75% for English, what is a satisfactory result, taking into account that flat unigram language model and relatively large word networks were used.

All experiments were done by setting the *word insertion penalty* in HTK *HVite* (p) to –60, which significantly

affects the word accuracy through varying the number of word insertions.

5.2 Speaker Independent Training and Evaluation

In order to evaluate the loss of accuracy due to speaker independent training the second experiment was performed using the second set with 10 speakers (C_10_500 and E_10_500) and these results were compared to the results of the first experiment. The model was built and tested 10 times. In each trial, utterances from one speaker were selected as a testing set and the remaining 90% of the set were used for training purpose, with the same parameters as in the first experiment, what is also known as *leave one out* or *cross validation* method. As such, it gives the worst case performance of the ASR system, when existing acoustic model built from a closed set of speakers is used for recognition of a newly enrolled subject without any adaptation.

Average results together with standard deviations and the maximum and minimum scores are shown in Table 4. For the Croatian language, the average word correctness score dropped to 66%, while for English to 44%. Although speaker independent models clearly impair recognition score in this experiment, such training is expected to outperform the speaker dependent training when built using a speech corpus of a proper size and sufficient number of speakers and when evaluated using the test data base with the unlimited number of speakers.

5.3 Grammar Influence

In the third experiment, the grammar influence was tested using the phone dialing sentences collected in the third set of the speech corpus (C_1_50_pd and E_1_50_pd). We have created a word network using a phone dialing grammar that strictly defines possible word-to-word transitions. It was used to test the same acoustic models trained in the first experiment. In this grammar, only names, surnames or phone numbers are allowed combined with a few dialing commands, as shown in Fig. 3.

Word networks were also much smaller for this case with approximately 25 words for both languages. The grammar used in this evaluation is abbreviated with *Gpd* in Table 4. The hundred percent accuracy was obtained for both languages, what shows that the grammar is particularly useful in applications that have a regular and strict word order in a sentence.

5.4 Cross-Word Transcription Influence

There are two ways of transcribing continuous speech. Word internal is the default method in this article and it is applied in all experiments discussed so far. In our fourth

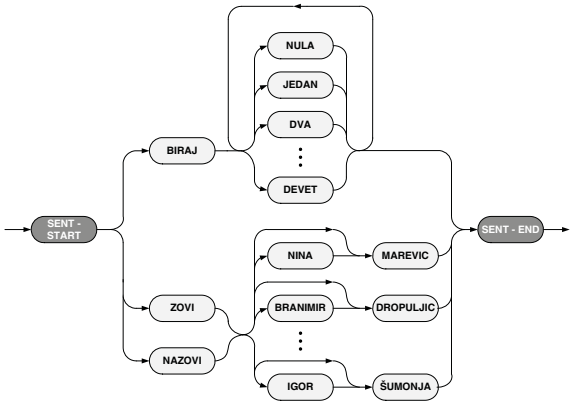

Fig. 3. Phone dialing grammar in Croatian language

experiment, we have performed a new training and evaluation analogous to the first experiment, but in this case we have used the cross-word transcription method. This means that words are transcribed by using a cross word triphones allowing for better modeling of concatenated words observed in the continuous speech pronunciation. Phonetic questions that we have been used for state tying are also important for cross-word models, since word concatenation might easily introduce new triphones not represented in the training set.

Slight improvement (+0.5%) offered by the cross-word transcription was observed for the Croatian model for word correctness, but otherwise, word accuracy for Croatian language as well as both scores for English language, are worse than with the default word-internal method. Probable reason for such a behavior is the limited training set size and significant increase of the complexity of word networks due to all possible additional transitions required by the cross-word triphones.

5.5 Croatian Transcription Rules Influence

Transcription rules (Table 2.) for the acoustic model of Croatian language were implemented within the developed ASR training application according to the described CLT algorithm. Influence of these rules on the recognition results was analyzed in the fifth experiment. In order to determine how often the CLT rules are applied, we have processed two word lists and counted number of words affected by CLT. For the first list comprised of 15000 most common words in Croatian language the percentage of modified words was 16.65%. This list was compiled by the Institute of Croatian Language and Linguistics, under the project of *Croatian Language Repository*. The second word list was derived from sentences of the C_1_500 data set. In this case, the number of word affected by CLT was 15.56%. Typically only one or few phonemes in such

words are actually affected (either by deletion, substitution or insertion).

In order to evaluate the influence of CLT; the same model was rebuilt as in the first experiment, but without using CLT rules. In this case, a simple one-to-one mapping between phonemes and graphemes was used. Syllabic *r* /ṛ/ and the diphthong /ie/ were excluded and monophone models were built from 30 phonemes directly corresponding to 30 graphemes of the Croatian alphabet. Only a small decrease of 0.2% (in both accuracy and correctness) can be observed in Table 4. when CLT rules are deactivated. Certainly, the trained triphone models are capable of capturing pronunciation changes due to the left and right context that are otherwise explicitly modeled through the CLT rules. Probable reason for relatively small difference is the fact that only a small percentage of phonemes in average are affected by CLT. A specially designed test dataset comprised of only the words affected by CLT is expected to exhibit greater difference.

Nevertheless, although this improvement is not substantial, it is useful since it can provide better results during online recognition, while only demanding additional engagement at initial phase when creating pronouncing dictionary and building models and word network.

Table 4. Testing results

Croatian Language				
Experiment	***Description***		***Word Corr. (%)***	***Word Acc. (%)***
Speaker Dependent	Tr: C_1_500 (90%), w-i, CLT Te: C_1_500 (10%), Guf		90.13	87.77
Speaker Indep. 10x(leave one out)	Tr: C_10_500 (9SP), w-i, CLT Te: C_10_500 (1SP), Guf	avg	65.56	61.66
		std	9.81	11.23
		max	71.77	68.92
		min	38.76	31.34
Grammar Influence	Tr: C_1_500 (90%), w-i, CLT Te: C_1_50_pd, Gpd		100.00	100.00
C-W Influence	Tr: C_1_500 (90%), c-w, CLT Te: C_1_500 (10%), Guf		90.56	85.19
Without CLT	Tr: C_1_500 (90%), w-i Te: C_1_500 (10%), Guf		89.91	87.55
English Language				
Experiment	***Description***		***Word Corr. (%)***	***Word Acc. (%)***
Speaker Dependent	Tr: E_1_500 (90%), w-i Te: E_1_500 (10%), Guf		75.14	66.18
Speaker Indep. 10x(leave one out)	Tr: E_10_500 (9SP), w-I Te: E_10_500 (1SP), Guf	avg	44.28	37.67
		std	7.89	8.78
		max	58.37	51.36
		min	31.89	21.78
Grammar Influence	Tr: E_1_500 (90%), w-i Te: E_1_50_pd, Gpd		100.00	100.00
C-W Influence	Tr: E_1_500 (90%), c-w Te: E_1_500 (10%), Guf		67.34	57.02

6 CONCLUSION AND FUTURE WORK

Results presented in the paper show that the acoustic model for a continuous speech ASR system for Croatian language can be easily developed with good quality results. Achieved word correctness was 90.13% for a speaker depended system and 65.56% for the speaker independent one, with a vocabulary size of 2403 words in both cases. Results for the Croatian model are actually better than the results for the English model that was trained and tested in an identical setup. This proves that HTK tool used in our experiments is indeed easily customized to any new language. Better result for the Croatian model, compared to the English one, probably has to do with the fact that Croatian language is phonetically simpler. It is therefore easier to train and model using a HMM. Another possible reason is the fact that only native Croatian speakers participated in the experiments. The difference was particularly observed for the case when speaker independent data set was used for training. It is expected that speakers will make fewer unnatural variations and mistakes when uttering their native language, what explains this difference. More detailed results are described in [10] and [16].

We have shown that application of the proposed algorithm for automatic transcription of Croatian words indeed improves the recognition score. We plan to upgrade this basic algorithm by enabling it to transcribe compound words, acronyms, numbers, dates and other words that are not transcribed in accordance to the standard rules of transcription in Croatian language.

In described experiments we have used a recorded speech corpus with read sentences using non-emotional speech. Further experiments will be carried out with a newly developed corpus recorded during actual stress training sessions, in order to verify the effects of emotional speech on the recognition accuracy. So far, our focus was on development and verification of the automated tool, using mockup databases. Nevertheless, we have shown that speaker dependent models can be built using limited data sets, what is especially important for targeted applications. The current version of the ASR system was already successfully applied for voice control of a robotic arm with two joints and for an automated voice entry of exam results at our University.

Additionally, future work includes development of statistical language models of Croatian language to improve current results that rely only on developed acoustic models. Such language models will be trained for specific applications like the emotional state estimator in stress inoculation training (SIT) and VR exposure therapy (VRET).

ACKNOWLEDGMENT

This work was supported by the Ministry of Science, Education and Sports of the Republic of Croatia under

project #036-0000000-2029. Authors would like to thank colleagues at the Dpt. of Electronic Systems and Information Processing for useful discussion and comments during preparation of the manuscript, as well as to anonymous reviewers for their valuable suggestion in preparing the final version.

REFERENCES

[1] L. E Baum and T. Petrie, "Statistical interface for probabilistic functions of finite state Markov chains," *The Annals of Mathematical Statistics*, vol. 37, pp. 1554-1563, 1966.

[2] J. K. Baker, "The dragon system – An overview," *IEEE Transactions on* Acoustics, Speech and Signal Processing, vol. 23, pp. 24-29, Feb. 1975.

[3] F. Jelinek, L. R. Bahl and R. L. Mercer, "Design of a linguistic statistical decoder for the recognition of continuous speech," *IEEE Transactions on Information Theory*, vol. 21, pp. 250-256, May. 1975.

[4] S. Martinčić-Ipšić and I. Ipšić, "Context-Dependent Acoustic Modeling of Croatian Speech," in *Proc. IS-LTC'06*, 2006, pp. 251-257.

[5] S. Martinčić-Ipšić, R.Slobodan and I. Ipšić, "Acoustic Modelling for Croatian Speech Recognition and Synthesis," *Informatica*, vol. 19, pp. 227-254, 2008.

[6] S. Martinčić-Ipšić, "Croatian Speech Recognition and Synthesis Based on Context-dependent Hidden Markov Model," PhD Thesis, in Croatian, University of Zagreb, Croatia, Nov. 2007.

[7] S. Dobrišek, F. Mihelič and N. Pavešić, "Speech Segmentation Aspects of Phone Transition Acoustical Modelling," *Lecture Notes in Computer Science*, vol. 1692, pp. 844, Jan. 1999.

[8] A. Zgank, Z. Kacic and B. Horvat, "Comparison of Acoustic Adaptation Methods in Multilingual Speech Recognition Environment," *Lecture notes in computer science*, vol. 2807, pp. 245-250, Feb. 2004.

[9] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, *The HTK Book (for HTK Version 3.3)*, 8th ed., Cambridge University Engineering Department, 2005.

[10] B. Dropuljić, "Development of Acoustic and Lexical model for Automatic speech recognition for Croatian Language," Diploma Thesis, in Croatian, University of Zagreb, Croatia, Jan. 2008.

[11] B. Schuller, G. Rigoll and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. ICASSP '03*, 2003, pp. II - 1-4.

[12] B. Schuller, G. Rigoll and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Proc. ICASSP '04*, 2004, pp. I - 577-580.

[13] T.S. Polzin, "Verbal and non-verbal cues in the communication of emotions," in *Proc. ICASSP '00*, 2000, pp. 2429-2432.

[14] D. Kukolja, S. Popović, B. Dropuljić, M. Horvat and K. Ćosić, "Real-time emotional state estimator for adaptive virtual reality stimulation," *Lecture Notes in Computer Science*, vol. 5638, pp. 175-184, Jul. 2009.

[15] S. Popović, M. Horvat, D. Kukolja, B. Dropuljić and K. Ćosić, "Stress inoculation training supported by physiology-driven adaptive virtual reality stimulation," *Studies in Health Technology and Informatics*, vol. 144, pp. 50-54, 2009.

[16] N. Marević, "Development of Acoustic and Lexical model for Automatic speech recognition for English Language," Diploma Thesis, in Croatian, University of Zagreb, Jan. 2008.

[17] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164-171, 1970.

[18] S. B. Davis, and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on* Acoustics, Speech and Signal Processing, vol. 28, pp. 357-366, Aug. 1980.

[19] International Phonetic Association, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, Cambridge University Press, Massachusetts, 1999.

[20] S. Babić, D. Brozović, I. Škarić and S. Težak, *Sounds and Forms of Croatian Literary Language, A Great Croatian Grammar*, in Croatian, Globus, Zagreb, 2007.

[21] B. Byrne, M. Finke, S. Khudanpur, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters and G. Zavaliagkos, "Pronunciation modeling for conversational speech recognition: A status report from WS97," in *Proc. ASRU '97, IEEE Workshop on*, 1997, pp. 26-33.

[22] D. Petrinović, *Digital speech processing*, lecture materials, in Croatian, Faculty of electrical engineering and computing, University of Zagreb, 2002.

[23] E. Barić, M. Lončarić, D. Malić, S. Pavešić, M. Peti, V. Zečević and M. Znika, *Croatian Grammar*, in Croatian, Školska knjiga Zagreb, 1997.

[24] J.C.Wells, *Longman Pronunciation Dictionary*, 1st ed., Harlow: Longman, 1990.

[25] J.A.C. Badenhorst, M.H.Davel, "Data requirements for speaker independent acoustic models," in *Proc. of the 19th Annual Symposium of the Pattern Recognition Association of South Africa*, Nov. 2008, pp. 147-152.

Branimir Dropuljić was born in Varaždin, Croatia in 1983. He received the Dipl.ing. degree in Electrical Engineering from University of Zagreb Faculty of Electrical Engineering and Computing in 2008. Since 2008, he has been research assistant at the Department of Electric Machines, Drives and Automation, Faculty of Electrical Engineering and Computing. He is doctoral student working on the scientific project "Adaptive control of virtual reality scenarios in therapy of post-traumatic stress disorder (PTSD)". His interests and main filed of research include emotional state estimation from vocal and psychophysiological responses.

Davor Petrinović was born in 1965 in Croatia. He received the Dipl.ing. degree in Electrical Engineering from University of Zagreb Faculty of Electrical Engineering in 1988 (today, Faculty of Electrical Engineering and Computing). He received M.Sc. and Dr.Sc. degree in the field of electrical engineering, in 1996. and 1999. respectively, from the same institution. He was appointed an Associate Professor in 2005, at the Department of Electronic Systems and Information Processing, Faculty of EE&C, Uni. Zagreb. He was a Fulbright post. doc. scholar in 2000/01 at SCL Laboratory, UC Santa Barbara, USA and a visiting researcher at Sound and Image Processing Lab, School of EE, KTH, Stockholm Sweden in 2005/06. His current research interests include speech and audio modeling, processing and coding.

AUTHORS' ADDRESSES

Branimir Dropuljić, Dipl.ing.
Department of Electric Machines, Drives and Automation,
Prof. Davor Petrinović, Ph.D.
Department of Electronic Systems and Information Processing,
Faculty of Electrical Engineering and Computing,
University of Zagreb,
Unska 3, HR-10000 Zagreb, Croatia
emails: branimir.dropuljic@fer.hr, davor.petrinovic@fer.hr

Received: 2010-01-21
Accepted: 2010-03-12