Whispered Speech Recognition using Hidden Markov Models and Support Vector Machines

Jovan Galić[1,2], Branislav Popović[3], Dragana Šumarac Pavlović[1]

[1]School of Electrical Engineering, University of Belgrade, Bulevar Kralja Aleksandra 73, 11120 Belgrade, Serbia

[2]Faculty of Electrical Engineering, University of Banja Luka, Patre 5, 78000 Banja Luka, Bosnia and Herzegovina

[3]University of Novi Sad, Faculty of Technical Sciences, Department of Power, Electronic and Telecommunication Engineering, Chair of Telecommunications and Signal Processing, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia

E-mails: jovan.galic@etf.unibl.org, bpopovic@uns.ac.rs, dsumarac@etf.rs

Abstract: Whisper is a specific mode of speech characterized by turbulent airflow at the glottis level. Despite an increased effort in speech perception, the intelligibility of whisper in human communication is very high. An enormous acoustic mismatch between normally phonated (neutral) and whispered speech is the main reason why modern Automatic Speech Recognition (ASR) systems have significant drop of performances when applied to whisper. In this paper, we present an analysis in recognition of whisper using 2 machine-learning techniques: Hidden Markov Models (HMM) and Support Vector Machines (SVM). The experiments are conducted in both Speaker Dependent (SD) and Speaker Independent (SI) fashion for Whi-Spe speech database. The best neutral-trained whisper recognition accuracy in SD fashion (83.36%) is obtained in SVM framework. At the same time, HMM-based recognition gave the highest recognition accuracy in SI fashion (87.42%). The results in recognition of neutral speech are given as well.

Keywords: Automatic Speech Recognition; Hidden Markov Models; Support Vector Machines; Whispered speech; Whi-Spe speech database

1 Introduction

Speech is the most natural and convenient form of interpersonal communication. According to modality, speech can generally be classified into 5 modes: whispered speech, soft speech, normally phonated speech (normal or neutral speech), loud speech and shouted speech [43]. Whisper is a specific mode of speech characterized by an absence of glottal vibrations and noisy excitation of the vocal tract. It is often used in daily life, especially over cellular phones.

Humans tend to whisper or generally lower their voice in an environment where normal speech is prohibited or inappropriate (e.g. in theater or reading room). An alternative way of communication is achieved with whisper if some confidential information should not be overheard. Whisper is sometimes used in criminal activities for hiding a speakers' identity. In addition to conscious production of whispered speech, it may also be phonated due to health issues, which appear after laryngitis or rhinitis [22].

In spite of the fact that vast improvements in speech technologies has been made in the last two decades, some disadvantages remain. The major drawback is the considerable performance degradation in adverse conditions, i.e., for speech that deviates significantly from the training data. Also, speech technologies are designed for recognition of the most commonly used mode of phonation, i.e. the neutral speech. In a range of speech modes from whisper to shouted, whispered speech has the most negative impact on the performance of Automatic Speech Recognition (ASR) systems [43]. Since whisper data are not generally available (or at least not in a sufficient amount), the greatest issue is confined to the automatic speech/speaker recognition in whispered mode, while training is done on normally phonated speech.

Classification technique based on Support Vector Machines (SVM) has shown good robustness in many speech recognition tasks, due to its operation principle based on finding the optimal separating hyper-plane that maximizes the margin between classes of the training data. The goal of this paper is to analyse the application of SVM in ASR systems for the recognition of *bimodal speech,* i.e., the neutral speech and whisper, and to compare the performance with the traditional HMM-based framework. Special attention is paid to whispered speech recognition improvement in the case where training is completed on utterances in normal phonation (N/W scenario). This study includes the analysis of recognition in both speaker dependent (SD) and speaker independent (SI) fashion. The recognition of isolated words (from a constrained lexicon) uttered in normal and whispered phonation in different train/test scenarios is considered.

The remainder of this paper is organized as follows. An overview of related works is briefly summarized in Section 2. In Section 3, the basic characteristics of whispered speech and its comparison with normal speech are given. Section 4 provides the ASR methodology based on Hidden Markov Models (HMM) and Support Vector Machines (SVM), whereas experimental preparation is described in Section 5. Experimental results, as well as their discussion, are presented in Section 6, for both the SD and SI recognition. Finally, the concluding remarks and the directions for future improvements are presented at the end of the paper.

2 Related Works

One of the earliest research studies in recognition of whispered speech was conducted for the Japanese language at the University of Nagoya [19]. The research has shown that the accuracy of whispered speech recognition can be effectively increased by using small amount of whispered speech for the adaptation of target speaker. The following studies were focused on the compensation of differences between neutral and whispered speech. Significant improvement for whisper speaker identification was obtained with frequency warping and score competition [9].

The generation of pseudo-whisper for efficient model adaptation based on Vector Taylor Series (VTS) algorithm was demonstrated in [13, 14]. Together with vocal tract length normalization and shift frequency transformation the Word Error Rate (WER), reduction from 27.7% to 17.5% (for open speaker scenario) was reported. The ASR system was speaker independent with lexicon constrained to 160 words.

High accurate detection of whisper-islands embedded within continuous neutral speech was achieved with linear prediction residual and entropy-based features [40, 41]. Whisper recognition based on deep neural networks and KALDI toolkit was investigated in [25]. Alternative techniques for recognition of normally phonated and whispered speech were examined using non-audible murmur microphone [1, 35], microphone arrays [42], throat microphone [21] and using camera for lip-reading and obtaining video features simultaneously with audio features [8].

The use of Teager energy cepstral coefficients with deep denoising autoencoder (DDA) has recently brought many benefits in speaker dependent (SD) neutral-trained whisper recognition [18]. Likewise, performances of speaker independent (SI) recognition of whispered speech have been significantly improved after adapting the acoustic model toward the DDA pseudo-whisper samples, compared to the model adaptation on an available small whisper set (for *UT-Vocal Effort II* speech corpus) [13, 15]. However, to the best of our knowledge, comparison of different speech recognition tools that include SVM in recognition of whispered speech was not reported. In this paper, comparison of HMM and SVM-based recognition of whispered speech is analyzed in both the SD and SI fashion.

Recently, using Teager energy cepstral coefficients has brought many benefits in speaker dependent neutral-trained whisper recognition [27]. Significant improvement in whisper recognition is achieved using cepstral coefficients with μ-law frequency warping [11].

Preceding papers related to the recognition of whispered speech from Whi-Spe database [26] were focused on the SD recognition. In [17], signal pre-processing procedure based on spectral whitening (so-called inverse filtering) improved neutral-trained whisper recognition. The comparison between different

normalization techniques was analyzed in [16]. The best results were obtained by using Cepstral Mean Normalization (CMN).

The research presented it this paper represents an important issue in Cognitive Infocommunications (CogInfoCom) [2]. By definition, CogInfoCom addresses the connection among research areas of infocommunications and cognitive sciences and their engineering applications. The goal of CogInfoCom is to provide a systematic view of how cognitive processes can co-evolve with infocommunications devices, and how humans may interact with the capabilities of artificially cognitive systems [3]. By the cognitive linguistic view, it has been stated that language represents an emergent cognitive capability, inseparably intertwined with the way in which we interact with the environment [7]. Several aspects of human-computer interaction (HCI) could be considered in CogInfoCom. One of those aspects is a negative impact of reduced resolution in HCI and multimodal interaction systems [3]. The problem of whispered speech recognition clearly addressees the issue.

Generally speaking, human-computer speech communication has been one of the most popular topics in CogInfoCom research area. In paper [31], the authors discuss including filled pauses and disfluent events into the training data for statistical language modeling, in order to improve speech recognition accuracy and robustness in the case of spontaneous speech. In [30], speech analysis has been conducted to verify the speaker authorization and measure the stress level within the air-ground voice communication, to improve voice communication in air traffic management security. In [29], a report is made on a set of perceptual experiments designed in order explore the human ability to identify emotional expressions presented through visual and auditory channels. A special emphasis is given to the cultural context, and in particular the language, and its influence to the study. In [24], the authors provide statistical analysis, examination and classification of features. Then they compare their discriminability in the case of read and spontaneous speech, for the task of automatic detection of depression by speech processing. In [37], automatic stress detection and prosodic phrasing approaches have been applied on pathological speech samples, in order to examine their discrimination capabilities in analyzing the samples from healthy and non-healthy individuals. A method used to collect video and audio recordings of people interacting with a simple robot interlocutor is proposed in [38]. In [34], a large-scale subjective study of phase importance in digital processing of speech is provided.

Although a novel contribution was presented in each study, implementation of a commercially available speaker independent recognizer for whispered speech remains an important issue that needs to be addressed in more details.

3 Whispered Speech Characteristics

Whisper represents a specific kind of speech, which is by its characteristics, nature and generating mechanism quite different from normal speech. The main characteristic of whisper is an absence of fundamental frequency and noisy excitation of the vocal tract. It was determined that formant frequencies for whispered vowels are substantially higher than for the normal voice [23]. The perceived pitch of whispered vowels was found to be very close to the second formant [36]. Compared to normally phonated speech, whisper has lower frame energy, longer durations of speech and silence, flatter long-term spectrum and lower sound pressure level (SPL) [43]. Despite of an increased effort in speech perception, the intelligibility of whisper is very high [33]. The auditory perception of emotional Chinese whispered speech has demonstrated that whispered speech can also carry some emotional information as the voiced one do [6]. On the other hand, non-linguistic information, such as age, sex or identity is hardly revealed in whisper.

In Figures 1 and 2, the waveform and the spectrogram of the short sentence in Serbian "Govor šapata." ("Whispered speech."), uttered in normally phonated and whispered speech, are depicted, respectively. The figures are supported with phonetic transcriptions. Because of the lack of sonority, a difference in amplitude levels between the two modes of speech can be observed. However, the spectrograms show that some parts of spectrum are well preserved in whisper, especially in the case of unvoiced consonants, such as fricative /š/ (/ʃ/ in IPA notation) and plosives /p/ and /t/. A similar shape of spectrum of vibrant /r/ in Serbian is observed. Moreover, the spectrogram shows that the harmonic structure of vowels is lost in the case of whisper.

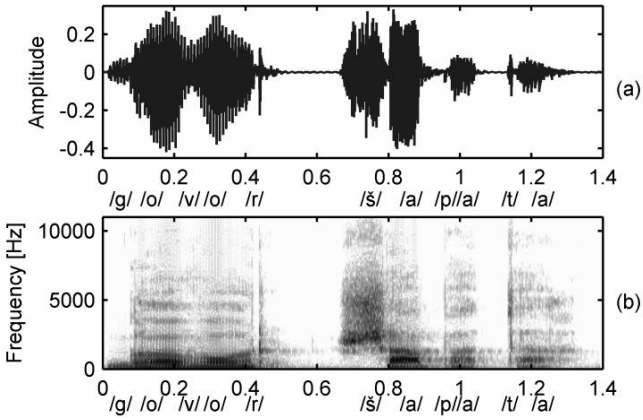

Figure 1

The waveform (a) and the spectrogram (b) of the short sentence "Govor šapata." (Whispered speech) uttered in normal phonation. The time in seconds is given on the abscissa.

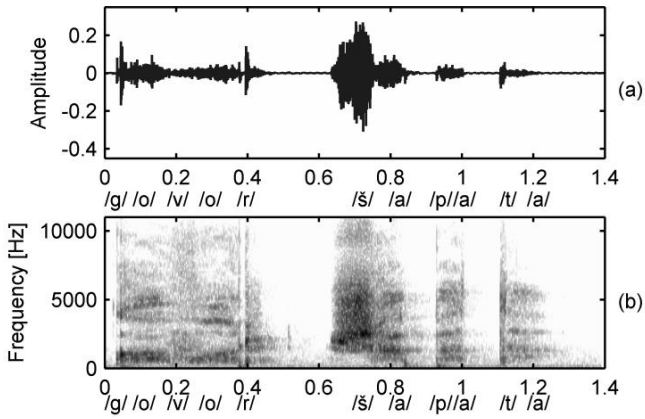

Figure 2

The waveform (a) and the spectrogram (b) of the short sentence "Govor šapata." (Whispered speech) uttered in whispered speech. The time in seconds is given on the abscissa.

4 Experimental Setup

4.1 The Whi-Spe Speech Database

The greatest issue in the utilization of whispered speech in ASR systems is the lack of an extensive and appropriate speech database. Therefore, a Whi-Spe speech database (abbreviation of *Whispered Speech*) was created for this research [26]. The database was recorded in quiet laboratory conditions, with a high-quality omni-directional microphone in mono technique. It was designed to have two parts: one that contains recordings of whispered words, and another one that contains recordings of the same words uttered in normal phonation. The corpus of 50 words spoken by 10 speakers (5 of them male and 5 female) was included in the database. Each speaker read all the words 10 times in both speech modes. Finally, the Whi-Spe corpus contained 5000 recorded words in normal speech and the same number of words recorded in whisper, or 2 hours in total. The speech data were digitized using a sampling frequency of 22050 Hz and 16 bits per sample, in Windows linear Pulse Code Modulation (PCM) *.wav* format.

During a recording session, each speaker read all the words continuously. The recording sessions were organized more than 10 times (with a pause of a few days between successive recordings) in order to collect a sufficient number of good quality representatives. The quality control of recordings found various types of errors. Some of them were related to an incorrect articulation or a wrong pronunciation, but most of them were related to the whispered speech. One of the

major problems of whispered recordings was insufficient signal level in relation to the ambient noise.

The specific details about the vocabulary of the Whi-Spe database, manual segmentation procedure, quality control and a labeling can be found in [26].

4.2 The Characteristics of the HMM-based ASR System

In ASR systems, the conventional technology is based on HMMs with Gaussian mixture models (GMMs). The most commonly used modeling units in isolated words recognition are phonemes independent from their context (monophones), phonemes dependent from their context (usually biphones or triphones), and the whole words. The greatest robustness in the case of experiments with the Whi-Spe database was achieved for the monophone models [10]. Therefore, models of phonemes independent from their context and Mel Frequency Cepstral Coefficients (MFCC) were used in this research.

For the extraction of feature vectors, the Hidden Markov Model Toolkit (HTK) software [39] was used. The Hamming window with pre-emphasis coefficient of 0.97 was used in order to obtain a feature vector. The window size was set to 24 ms, with the frame shift of 8 ms. In filterbanks, the power cepstrum was used rather than the magnitude. Each frame was represented with 39 coefficients, i.e., 13 cepstral coefficients (including the energy), along with their first and second order time derivatives. The coefficients were normalized with cepstral mean of each utterance.

The ASR system backend was based on HMM models with output probabilities modeled using the continuous density GMMs and diagonal covariance matrices. Each monophone model was represented with 5 states in total (3 emitting states) with strictly left to right topology and without skips. Each word from the Whi-Spe database was transcribed manually. The number of training cycles in embedded re-estimation was set to 5 and the variance floor for Gaussian probability density functions was set to 1%. The number of mixture components was 8 for the SD recognizer and 32 for the SI recognizer and it was gradually increased. In the testing phase, the Viterbi algorithm was applied in order to determine the most probable state sequence. The phone level transcription was performed with 32 monophone units - 30 monophones corresponding to 30 letters in the Serbian alphabet, the phoneme schwa and the silence. The phoneme schwa is marked when /r/ is found in a consonant environment, whereas the model of silence was appended at the start and at the end of each utterance. The ASR system developed in this study was completely implemented using HTK. The generation of the script and configuration files, as well as the files for model initialization and phonetic transcription was automated using MATLAB. MATLAB was also utilized for logging the ASR system performance results with an evaluation in HTK.

4.3 The Characteristics of the SVM-based ASR System

The SVM is a relatively simple and efficient machine-learning algorithm, which is widely used for pattern recognition and classification problems, especially under the condition of data-sparsity. The underlying concept behind the SVM is the structural risk minimization. It is supervised classification algorithm with good generalization properties when number of training patterns is limited. For that reason, we examined the performance of SVM in whisper recognition.

SVM was initially introduced for classifying linearly separable classes of objects. The separation of classes into 2 categories is obtained by using an n-dimensional hyper-plane that maximizes the margin between classes. In most real-world classification problems, classes are not linearly separable. In that case, non-linear feature-vector transformation is performed in order to map into high-dimensional feature space, in which linear separation of classes is expected. A function used for mapping is called kernel. Some common kernels include:

- Radial basis function kernel (adjustable parameter γ)

$$K(x_1, x_2)=\exp\left(-\frac{\left\|x_1-x_2\right\|^2}{2\sigma^2}\right);\ \gamma=\frac{1}{2\sigma^2}$$

- Polynomial kernel (adjustable parameters are the slope α, the constant term c and the polynomial degree d)

$$K(x_1, x_2)=(\alpha x_1^T x_2+c)^d;$$

- Linear kernel (adjustable parameter c)

$$K(x_1, x_2)=x_1^T x_2+c$$

- Hyperbolic Tangent (Sigmoid) kernel (adjustable parameter are α and the constant term c)

$$K(x_1, x_2)=\tanh(\alpha x_1^T x_2+c)$$

There are typically 2 approaches to solve non-binary classification problems. The first approach includes the comparison of each class against all the others (1-vs-all), whereas the second approach confronts each class against all the other classes separately (1-vs-1). In this study, the classifier with 1-vs-all comparison strategy was used because of better performance in multiclass HMM-SVM recognition of low-SNR isolated word utterances [5].

The fact that SVM is a static classifier is the main shortcoming for its widespread application in state-of-the-art ASR systems. Some hybrid SVM/HMM systems were developed in order to overcome that limitation [32].

Speech signal (being a stationary) should be analyzed on a short-time basis, in which it is assumed to be quasi-stationary. Typically, it is divided into a number of overlapping frames (usually Hamming windows) and feature vector is computed to represent each frame. The size of the analysis (w_s) is usually between 20 and 30 ms, with the frame shift f_p (time period between consecutive frames) between 10 and 15 ms. Therefore, utterances of different durations have unequal number of feature vectors. Two most common alternatives for making fixed number of frame windows for SVM classifier are [12]:

1) Variable window size - the window size is chosen to be proportional to the frame period ($w_s = Kf_p$), with overlapping factor *K* being constant for all utterances;
2) Fixed window size - the window length is fixed, but the overlapping factor is dynamically selected.

Both described procedures lead to loss of information, especially in the case of long speech utterances. In this paper, segmentation based on variable window size and constant *K* is chosen because of better performance in isolated words recognition using SVM technique [12]. The optimal number of windows per utterance depends on the related speech database and the corresponding lexicon. A heuristic search was made for SVM-based recognition of Spanish digits and was found to be 13 [12]. Therefore, in our initial experiments, we used utterance segmentation on 13 overlapping windows.

The MFCC speech parameterization is performed by using static features (13, including energy) along with the first and second order time derivatives (39 in total) and cepstral mean normalization. Finally, each utterance is represented with a vector of 507 coefficients (13x39), and that vectors are inputs for the SVM recognizer. Speech recognizer was developed in Python software package (version 3.6).

5 Results and Discussion

This section is organized as follows. The performances of the recognizer based on HMM framework are presented in subsection 5.1, while in subsection 5.2, the performances based on SVM framework are given. The experiments are conducted in both the SD and SI fashion, in 2 train/test scenarios:

1) N/N - the ASR system is trained on neutral speech and tested using the speech of the same mode. This scenario is marked as *matched*.
2) N/W - the ASR system is trained on neutral speech and tested against the speech of the opposite mode. This scenario is marked as *mismatched*.

In order to provide more reliable evaluation of the performance, cross-validation is needed. For each speaker, 1000 utterances (500 in neutral and 500 in whisper mode) are available. Word recognition accuracy is presented as a metric for evaluating the performance of the recognizer.

In the SD case, the accuracy is calculated according to the following procedure. In matched conditions, available utterances are divided in the train and test set. The train set contains 80% of utterances evenly distributed between words. Remaining 100 utterances are exploited in the test set. The recognizer displays the percentage of correctly recognized utterances. For example, if N denotes total number of analyzed utterances and E denotes the number of incorrectly recognized utterances, accuracy percentage is calculated in the following way:

$$accuracy = \frac{N-E}{N} \times 100[\%] \tag{1}$$

Train and test sets are rotated in 5-fold cross-validation. The accuracy for an examined speaker is calculated by averaging 5 results from cross-validation. Finally, the average recognition accuracy is computed as arithmetic mean of accuracies from all speakers. The procedure is the same for mismatched conditions, noting that the test set contains all available utterances in the opposite speech mode. In the train set, equal number of utterances is utilized in both matched and mismatched conditions.

In the SI case, all utterances from the examined speaker (for the respective mode) are given in the test set, whereas the utterances from the other 9 speakers in neutral speech are given in the train set (full dataset training with leave-one-speaker-out cross-validation). Again, the accuracy is averaged across different speakers.

5.1 HMM Framework

In the case of ASR systems based on continuous density HMMs, it is essential to provide good initial estimates of the HMM parameters, so that the Baum-Welch re-estimation algorithm could reach the global maximum of the likelihood function. If segmented data are available, the *k-means* algorithm can be used for calculation of the initial parameters, i.e., mean vectors and covariance matrices [39]. Additionally, instead of using a fixed number of states per each monophone model, a noticeable gain in robustness can be achieved with a variable number, proportional to the phoneme duration. The number of HMM states per model, proportional to the average duration of all the instances of the corresponding phone in the training database is proposed in [20], for all the phonemes in Serbian.

In this research, the number of states per model (two of which were non-emitting) for each monophone model is presented in Table 1.

Table 1
The number of states per monophone model. It should be noted that Serbian and IPA notations differ for the following consonants: ʃ (š), ɦ (x), ʒ (ž), ʦ (c), ʨ (ć), ʧ (č), ʣ (đ), ʤ (dž), ɲ (nj) and ʎ (lj).

Monophone	Number of states
/a/, /e/, /i/, /o/, /u/, /b/, /p/, /d/, /t/, /g/, /k/, /ʦ/, /ʧ/, /ʨ/, /ʤ/, /ʣ/, /s/, /ʃ/, /z/, /ʒ/, /f/, /ɦ/, /m/, /n/, /ɲ/	6
/j/, /l/, /ʎ/, /v/	5
/r/, /ə/	4
silence	3

Besides the recognition with flat-start initialization (in which models are initialized with the global mean and variance), the contribution of different initial estimates to the performance of the ASR system is analyzed as well. The parameters of the initial monophone models are obtained by using a small part of the database (10% of utterances in normal phonation) annotated with:

- Manual annotation;
- Automatic annotation with the forced alignment implemented in the HTK;
- Automatic annotation with the recognizer for the Serbian language based on the Kaldi speech recognition toolkit [28].

The manual annotation was done in the software package PRAAT [4]. For each word and each speaker in normally phonated speech, a phonetic expert labeled the start and the end time in one utterance (of total 10), by inspecting the time waveform and the spectrogram.

The HTK tool HVite can be used in automatic annotation systems and it operates in the so-called *forced alignment* mode. In this case, the recognition network is constructed from a word level transcription and a dictionary, instead of a task level word network (the default mode).

The Kaldi speech recognition toolkit can also be used in automatic annotation systems. The recognition models were trained using the Whi-Spe training data in 3 separate phases, i.e., the mono phase, the first and the second triphone phase, each with a different number of states (regression three leaves) and Gaussians. Each phase was initialized using the alignments from the previous phase. During the mono phase, 1000 Gaussians were employed. During the first triphone phase, 1800 states and 9000 Gaussians were used. During the second triphone phase, the system comprised 3000 leaves and 25000 Gaussians. The final model was applied to obtain the alignments used in order to calculate the initial parameters for flat-start ASR system.

The recognition results are depicted in Figure 3(a) and 3(b) for normal and whispered speech recognition, respectively. Depicted Standard Errors (SE) show standard deviation between different recognition systems divided by the square root of the sample size. For visual comparison of accuracies, an important part of each bar graph is emphasized (higher than 90% in matched and 40% mismatched scenario).

As one can see in Figure 3(a), compared to the recognition with flat-start initialization, the recognition of normal speech in the SD fashion (SD bars in Figure 3(a)), was improved for at least half a percent, regardless of the method of initialization. Because of the "ceiling effect" (accuracies are higher than 99.60%), it is hard to infer which manner of initialization gave the best performance. In the SI fashion (SI bars in Figure 3(a)), the annotation with the recognizer based on forced alignment gave a noticeable improvement (1.16%, absolute). The contribution of manual annotation and KALDI was with absolute increment of 0.8% (approximately).

The results presented in Figure 3(b) show that the recognition of whispered speech was improved for each manner of annotation, in both SD and SI fashion. Compared to the SD recognition with flat-start initialization (accuracy 73.42%), the greatest improvement was achieved with the manually annotated model initialization (accuracy 81.38%). However, in SI fashion (SI bars in Figure 3(b)), the greatest improvement was again achieved by using HTK, giving an accuracy of 87.42% (absolute increment 5.30%). The experiments showed marginal difference in terms of the accuracy between manual and KALDI annotation.

5.2 SVM Framework

As noted in 4.3, the most commonly used kernels in SVMs are RBF, sigmoid, linear and polynomial. However, each function that satisfies necessary properties (Mercer's theorem) can be used as a kernel. In this study, we examined performances of the recognizer for 4 mentioned kernels. The recognition results (average accuracy with SE) are depicted in Figures 4(a) and 4(b) for normal and whispered speech recognition, respectively.

As can be seen in Figure 4(a), for the recognition of normal speech, compared to HMM-based recognizer, there was a marginal decrease of the performance in the SD fashion, with maximum accuracy for linear kernel (99.26%). Still, the drop of the performance in the SI fashion was noticeable, with the best accuracy for polynomial kernel (95.93%).

The obtained results depicted in Figure 4(b) show that the highest accuracy in whisper recognition in SD fashion is with the sigmoid kernel (81.82%). However, the difference compared to the linear kernel is not significant. The use of the RBF kernel resulted in notably lower performance, whereas the polynomial kernel was practically useless. The best performance in the SI fashion was achieved using the

RBF kernel (75.29%, SI bars in Figure 4(b)), but it was far less successful than the HMM-based recognizer (SI bars in Figure 3(b)).

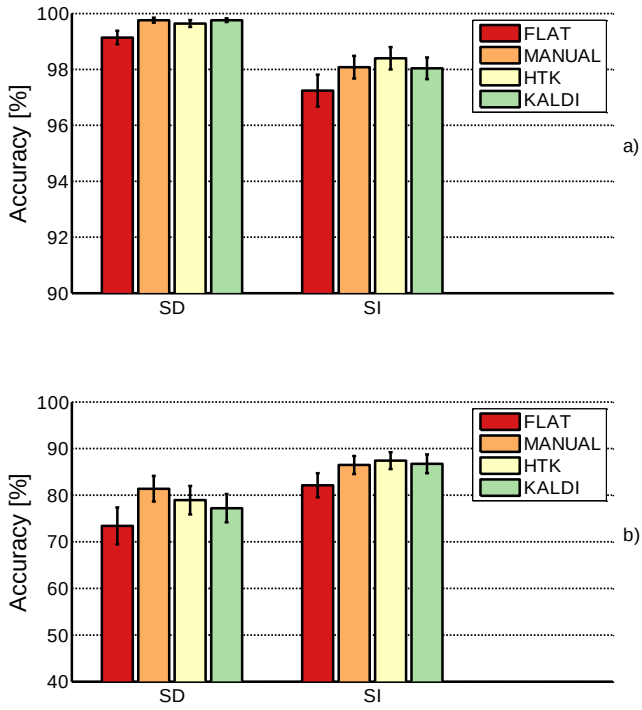

Figure 3

The HMM-based recognition accuracy with standard error (SE) in recognition of normal speech (a) and whisper (b) in speaker dependent (SD) and speaker independent (SI) fashion, in dependence of manner of annotation: FLAT - flat start; MANUAL - manual annotation; HTK - automatic annotation in HTK; KALDI - automatic annotation in KALDI

As already mentioned in subsection 4.3, we performed segmentation on 13 overlapping windows. The range of the number of phones per word in Whi-Spe speech database is from 3 to 13, whereas the average number is 5.58 (the longest words are very rare). Using 13 frames per word gives an average of one frame per phone in the longest word, while in short words there are two or three frames per phone.

We also tested performance on finer temporal and spectral resolution by using more than 13 windows per utterance in range from 13 to 19. The results are depicted in Figures 5 and 6 for recognition in the SD and SI fashion, respectively. Kernel with the best performance in the case with 13 windows was used.

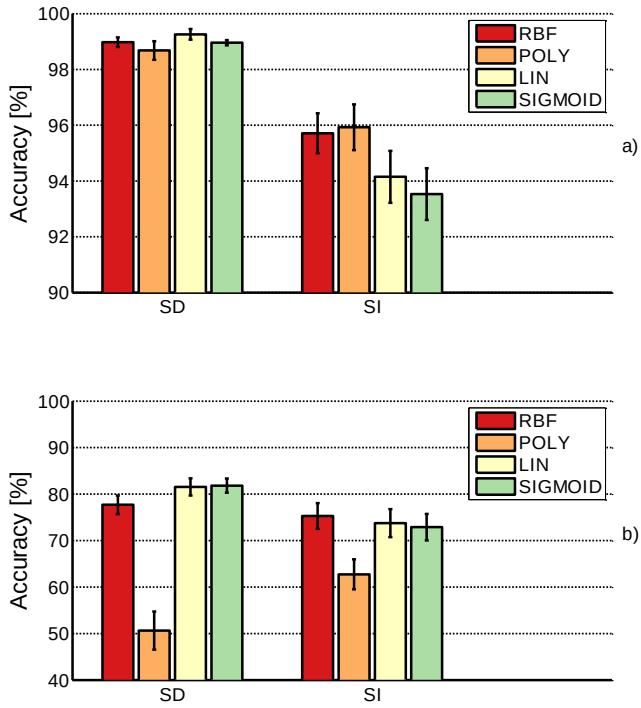

Figure 4

The SVM-based recognition accuracy with standard error (SE) in recognition of normal speech (a) and whisper (b), in speaker dependent (SD) and speaker independent (SI) fashion, in dependence of kernel

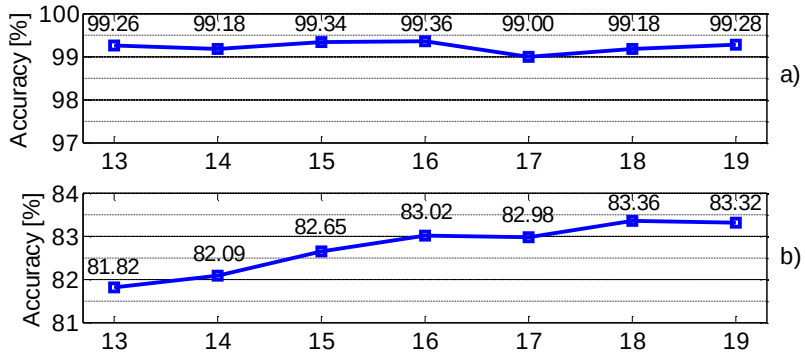

Figure 5

The SVM-based recognition accuracy in recognition of normal speech (a) and whisper (b) in speaker dependent (SD) fashion, in dependence of a number of windows

The results depicted in Figure 5(a) show that the change in a number of windows in the SD case give no statistically significant improvement in normal speech recognition. Yet, whisper recognition is improved for the additional 1.54% in the

case where utterances are segmented into 18 overlapping windows (see Figure 5(b)) with the average recognition accuracy of 83.36%.

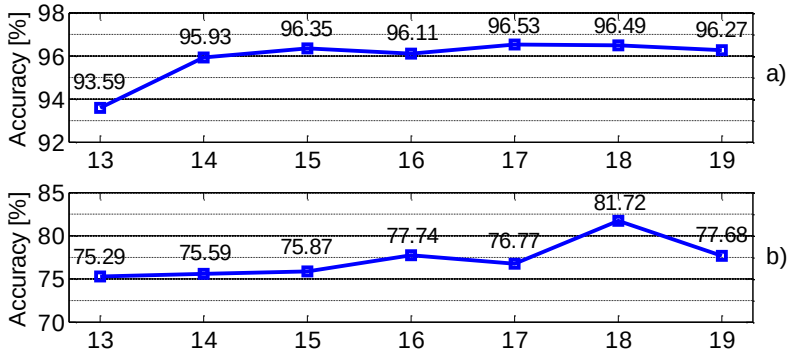

Figure 6

The SVM-based recognition accuracy in recognition of normal speech (a) and whisper (b) in speaker independent (SI) fashion, in dependence of a number of windows

The analysis in the SI fashion showed that improvement in normal speech recognition is about 3% (absolute) for 17 and 18 windows. Once again, the optimal number of windows in whisper recognition is 18, with an increment in recognition of 6.5%, approximately.

Conclusions

The motivation behind the research study presented in this paper is the growing need to raise human-computer speech communication to a higher level, which includes speech produced in phonation other than normal. Whispering is a useful mode of speech if someone does not want to be overheard, but in some cases it is unavoidable (damaged vocal tract, health issues, etc.).

The recognition of whispered speech with a satisfactory success (independent from speaker) is a serious challenge faced by speech technology scientists today. State-of-the-art ASR systems deal with this problem only in a restricted domain, with a constrained lexicon. The static nature of the SVM classifier is the main reason for good recognition accuracy only in the SD fashion, because the manner of utterance segmentation into fixed number of frames may lead to drop of useful speech information. The traditional HMM-based approach has given much better results in the SI fashion.

Since SVM classifier has better discrimination capabilities compared to the HMM, we hope that developing hybrid SVM/HMM ASR system may give better results in neutral-trained whisper recognition. This is the subject of our future work.

Acknowledgement

This work is partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under grants OI178027,

TR32032, and TR32035, EUREKA project DANSPLAT, "A Platform for the Applications of Speech Technologies on Smartphones for the Languages of the Danube Region", id E! 9944, and the Provincial Secretariat for Higher Education and Scientific Research, within the project "Central Audio-Library of the University of Novi Sad", No. 114-451-2570/2016-02.

The authors would also like to thank Professor Vlado Delić and Nikša Jakovljević for their kindness and help with some of the preliminary experiments for this work, which helped improve the quality of the paper.

References

[1] D. Babani, T. Toda, H. Saruwatari, K. Shikano, "Acoustic Model Training for Non-Audible Murmur Recognition using Transformed Normal Speech Data," Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP) Prague, Czech Republic, 2011, pp. 5224-5227

[2] P. Baranyi, A. Csapo, G. Sallai, "Cognitive Infocommunications," Springer Book, 2015

[3] P. Baranyi, A. Csapo, "Definition and Synergies of Cognitive Infocommunications," *Acta Polytechnica Hungarica*, Vol. 9, No. 1, pp. 67-83, 2012

[4] P. Boersma, D. Weenink,"Praat: Doing Phonetics by Computer [Computer program]," Version 5.3.51, retrieved 2 June 2013. Available from: http://www.praat.org/

[5] J. Bernal-Chaves, C. Peleaz-Moreno, A. Gallardo-Antolin, F. Diaz-de-Maria, "Multiclass SVM-based Isolated-Digit Recognition using a HMM-Guided Segmentation," Proceedings of Non-linaer Speech Processing, Barcelona, 2005, pp. 137-144

[6] G. Chenghui, Z. Heming, Z. Wei, W. Min, "A Preliminary Study on Emotions of Chinese Whispered Speech," Proceedings of International Forum on Computer Science-Technology and Applications (IFCSTA) Vol. 2, Chongqing, China, 2009, pp. 429-433

[7] W. Croft, D. A. Cruse, "Cognitive Linguistics," Cambridge University Press, 2004

[8] X. Fan, C. Busso, J. Hansen,"Audio-Visual Isolated Digit Recognition for Whispered Speech," Proceedings of European Signal Processing Conference (EUSIPCO) Barcelona, Spain, 2011, pp. 1500-1503

[9] X. Fan, J. H. L. Hansen, "Speaker Identification for Whispered Speech Based on Frequency Warping and Score Competition," Proceedings of Interspeech 2008, Brisbane, Australia, 2008, Vol. 1, pp. 1313-1316

[10] J. Galić, S. Jovičić, Đ. Grozdić, B. Marković, "HTK-based Recognition of Whispered Speech," Proceedings of International Conference on Speech and Computer SPECOM, Novi Sad, Serbia, 2014, pp. 251-258

[11] J. Galić, S. Jovičić, V. Delić, B. Marković, D. Šumarac Pavlović, Đ. Grozdić, "HMM-based Whisper Recognition Using μ-law Frequency Warping," accepted for publication in SPIIRAS Proceedings Journal, 2018

[12] J. M. Garcia-Cabellos, C. Peleaz-Moreno, A. Gallardo-Antolin, F. Perez-Cruz, F. Diaz-de-Maria, "SVM classifiers for ASR: A discussion about paramterization" Proceedings of 12th European Signal Processing Conference, 2004, pp. 2067-2070

[13] S. Ghaffarzadegan, H. Boril, J. H. L. Hansen, "Generative Modeling of Pseudo-Whisper for Robust Whispered Speech Recognition,"*IEEE/ACM Transactions on Speech and Language Processing*, Vol. 24, No. 10, pp. 1705-1720, 2016

[14] S. Ghaffarzadegan, H. Boril, J. H. L. Hansen, "Model and Feature Based Compensation for Whispered Speech Recognition," Proceedings of Interspeech 2014, Singapore, 2014, pp. 2420-2424

[15] S. Ghaffarzadegan, H. Boril, J. H. L. Hansen, "UT-Vocal Efort II: Analysis and Constrained-Lexicon Recognition of Whispered Speech," Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP) Florence, Italy, 2014, pp. 2544-2548

[16] Đ. Grozdić, S. Jovičić, D. Šumarac Pavlović, J. Galić, B. Marković, "Comparison of Cepstral Normalization Techniques in Whispered Speech Recognition," *Advances in Electrical and Computer Engineering (AECE) Journal*, Vol. 17, No. 1, pp. 21-26, 2017

[17] Đ. Grozdić, S. Jovičić, J. Galić, and B. Marković, "Application of Inverse Filtering in Enhancement of Whispered Speech," Proceedings of Neural Network Applications in Electrical Engineering (NEUREL) Belgrade, Serbia, 2014, pp. 157-162

[18] Đ. Grozdić, S. T. Jovičić, M. Subotić, "Whispered Speech Recognition using Deep Denoising Autoencoder," *Engineering Applications of Artificial Intelligence*, Vol. 59, pp. 15-22, 2017

[19] T. Ito, K. Takeda, F. Itakura, "Analysis and Recognition of Whispered Speech," *Speech Communication*, Vol. 45, pp. 129-152, 2005

[20] N. Jakovljević, "An Application of Sparse Representation in Gaussian Mixture Models used in Speech Recognition Task," PhD. thesis, University of Novi Sad, Faculty of Technical Sciences, 2013

[21] S. Jou, T. Schultz, E. Waibel, "Adaptation for Soft Whisper Recognition Using a Throat Microphone," Proceedings of International Conference on Spoken Language Processing (ICSLP) Jeju Island, Korea, 2004, pp. 1493-1496

[22] S. T. Jovičić, Z. M. Šarić, "Acoustic Analysis of Consonants in Whispered Speech," *Journal of Voice*, Vol. 22, No. 3, pp. 263-274, 2008

[23] S. T. Jovičić, "Formant Feature Differences between Whispered and Voiced Sustained Vowels," *Acta Acustica*, Vol. 84, No. 4, pp. 739-743, 1998

[24] G. Kiss, K Vicsi, "Comparison of Read and Spontaneous Speech in Case of Automatic Detection of Depression," Proceedings of CogInfoComm, pp. 213-218, 2017

[25] P. Kozierski, T. Sadalla, S. Drgas, A. Dąbrowski, D. Horla, "Kaldi Toolkit in Polish Whispery Speech Recognition," *Przeglad Elektrotechniczny*, pp. 301-304, 2016

[26] B. Marković, S. Jovičić, J. Galić, Đ. Grozdić, "Whispered Speech Database: Design, Processing and Application," Proceedings of 16th International Conference TSD, Pilsen, Czech Republic, 2013, pp. 591-598

[27] B. Marković, J. Galić, M. Mijić, "Application of Teager Energy Operator on Linear and Mel Scales for Whispered Speech Recognition", *Archives of Acoustics*, Vol. 43, No. 1, pp. 3-9, 2018

[28] B. Popović, E. Pakoci, S. Ostrogonac, D. Pekar: "Large Vocabulary Continuous Speech Recognition for Serbian Using the Kaldi Toolkit," Proceedings of 10th DOGS, Digital Speech and Image Processing, Novi Sad, Serbia, 2014, pp. 31-34

[29] M. T. Riviello, A. Esposito, "A Cross-Cultural Study on the Effectiveness of Visual and Vocal Channels in Transmitting Dynamic Emotional Information," *Acta Polytechnica Hungarica*, Vol. 9, No. 1, pp. 157-170, 2012

[30] M. Rusko, M. Finke, "Using Speech Analysis in Voice Communication: A New approach to improve Air Traffic Management Security," Proceedings of CogInfoComm, pp. 181-186, 2016

[31] J. Staš, D. Hladek, J. Juhar, "Adding Filled Pauses and Disfluent Events into Language Models for Speech Recognition," Proceedings of CogInfoCom, pp. 133-136, 2016

[32] Zhi-yi Qu, Yu Liu, Li-hong Zhang, Ming-xin Shao, "A Speech Recognition System Based on a Hybrid HMM/SVM Architecture," First International Conference on Innovative Computing, Information and Control (ICICIC) Beijing, China, 2006, pp. 100-104

[33] V. Tartter, "Identifiability of Vowels and Speakers from Whispered Syllables," *Perception & Psychophysics*, Vol. 49, No. 4, pp. 365-372, 1991

[34] L. Tesic, B. Bondzulic, M. Andric, B. Pavlovic, "An Experimental Study on the Phase Importance in Digital Processing of Speech Signal," *Acta Polytechnica Hungarica*, Vol. 14, No. 8, pp. 197-213, 2017

[35] T. Toda, K. Nakamura, T. Nagai, T. Kaino, Y. Nakajima, K. Shikano, "Technologies for Processing Body-conducted Speech Detected with Non-Audible Murmur Microphone," Proceedings of Interspeech 2009, Brighton, UK, 2009, pp. 632-635

[36] I. B. Thomas, "Perceived Pitch of Whispered Vowels," *Journal of Acoustical Society of America*, Vol. 46, No. 2, pp. 468-470, 1969

[37] M. Tündik, G. Kiss, D. Sztahó, G. Szaszák, "Assessment of Pathological Speech Prosody Based on Automatic Stress Detection and Phrasing Approaches," Proceedings of CogInfoComm, pp. 67-72, 2017

[38] B. Vaughan, J. G. Han, E. Gilmartin, N. Campbell, "Designing and Implementing a Platform for Collecting Multi-Modal Data of Human-Robot Interaction," *Acta Polytechnica Hungarica*, Vol. 9(1), pp. 7-17, 2012

[39] S. Young *et al*., "The HTK Book (for HTK Version 3.4)", Cambridge University Engineering Department, 2006 [Online] Available:http://speech.ee.ntu.edu.tw/homework/DSP_HW2-1/htkbook.pdf

[40] C. Zhang, J. H. L. Hansen, "Whisper-Island Detection Based on Unsupervised Segmentation with Entropy-based Speech Feature Processing," *IEEE Transactions on Audio Speech and Language Processing*, Vol. 19, No. 4, pp. 883-894, 2011

[41] C. Zhang, J. H. L. Hansen, "Advancements in Whisper-Island Detection using the Linear Predictive Residual," Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP) Dallas, USA, 2010, pp. 5170-5173

[42] C. Zhang, T. Yu, J. H. L. Hansen, "Microphone Array Processing for Distance Speech Capture: A Probe Study on Whisper Speech Detection," Proceedings of the Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, USA, 2010, pp. 1707-1710

[43] C. Zhang, J. H. L. Hansen, "Analysis and Classification of Speech Mode: Whisper through Shouted," Proceedings of Interspeech 2007, Antwerp, Belgium, 2007, pp. 2289-2292