Influence of the perceptual speech quality on the performance of the text-independent speaker recognition system

Robert Blatnik, Gorazd Kandus, and Tomaž Šef

***Abstract*—**In the following paper we examine the influence of the perceptual speech quality on the performance of the text-independent automated speaker recognition system (ASRS). The perceptual speech quality was objectively measured using Perceptual Evaluation of Speech Quality method (PESQ). The speech quality was degraded under various conditions as imposed in Voice over Wireless Local Area Network (VoWLAN), GSM and PSTN telephony. The ASRS error rates of this evaluation are presented by means of detection error tradeoff (DET) curves. The results show the correlations between PESQ MOS and ASRS equal error rate (EER) and promise the objective speech quality measurements can be used for the prediction of ASRS performance.

***Keywords*—** DET, MOS, PESQ, Speaker Recognition System, Speech quality.

I. Introduction

SPEECH degradations as imposed by various telephone networks have been proven to have large effects on the performance of the automated speaker recognition systems (ASRS) [1]. Moreover, the employment of various handsets, codecs and recording devices influences the performance of an ASRS. Performance degradation due to so-called *channel variability* has been already demonstrated during the past few evaluations conduced by the National Institute of Standards and Technology [2]. However, by the knowledge of the authors, there has not been substantial investigation of the correlations between ASRS error rates and measured speech quality of various transmission channels. The challenge is weather the perceptual quality can be used as a measure for predicting the error rates of ASRS.

Speech, as the medium of human communication conveys many types of information. Beside the message encoded in the language, the speaker also shares the information about its emotional and social state, health and other personal identifying characteristics such are: gender, age, dialect, voice, range of pitch, loudness and others [3]. Human voice combines physiological and behavioral characteristics of a certain speaker, which make it possible to distinguish one speaker from another. The characteristics of a certain speaker can be extracted and measured, which enables the automated speaker recognition system (ASRS) to decide whether two given speech recordings belong to the same speaker [4].

Any ASRS inevitably fail in certain amount of decisions which is commonly defined as the error rate. Error rates in ASRS occur due to changes in health, emotional state, age and other sources of variability of human voice. The fact that the same speaker recorded over different telephone networks, handsets or microphones sound differently is commonly referred as channel variability. As the channel variability is affecting ASRS performance, different telephone networks comprise different distortions, errors, noises, filtering, delay, jitter and others, commonly referred as the telephone-speech quality [5]. Moreover, the effect of noise-in-speech on the performance of a speaker recognition and speech recognition systems remains a challenging issue in the current research [6]. The telephony-speech quality can be evaluated subjectively by the listeners [7], or it can be objectively measured using Perceptual Evaluation of Speech Quality method (PESQ) [8].

As the main task of the ASRS is the correct decision in the process of identity verification of a certain speaker and we are not primarily interested in the transmitted message itself, on the other hand, the main attribute of the speech quality in the telephony is the intelligibility of the speech, and we are not primarily interested in the identity of the speaker.

The evaluations of ASRS usually require large amounts of speech recorded over various channels and conditions, extensive testing and analysis of such systems [2].

In the following paper we present an experimental evaluation of the ASRS performance and its relationship to the degradations of speech recordings transmitted over VoIP in wireless local area networks (VoWLAN), mobile telephony (GSM) and landline analogue telephony (PSTN). The speech quality degradations were objectively measured using PESQ method. The analysis show the correlations between PESQ mean option score (MOS) and ASRS error rates. The results of the experiment promise the objective speech quality measurements could be effectively used in the prediction of ASRS performance.

The reminder of this paper is organized as follows. After speech quality assessment methods presented in section 2, we introduce speaker recognition basics in section 3. We continue with short description of the ASRS performance measures in

This work is partially founded by the European Commission, the Swiss Federal Office for Education and Science, Slovenian Ministry of Higher Education, Science and Technology and Prevent Group, Slovenj Gradec.

section 4. The evaluation framework with ASRS experimental setup and PESQ speech quality test-bed for GSM, PSTN and VoWLAN is presented in section 5. In section 6 we present and discuss the results of ASRS evaluations and correlations with PESQ MOS. Finally, we conclude the paper in section 7.

II. Speech quality assessment

A. Speech quality assessment methods

Speech quality assessment methods could be generally divided into two main groups: subjective methods carried out by human listeners who perceptually evaluate the quality of speech under judgment and objective methods for speech quality assessment, carried out generally by machines [9]. The speech quality can be assessed from user perspective "perceived – subjective" or "objectively measured" with the parameters like delay, jitter and packet loss.

The traditional method for subject measurement of speech quality is to calculate a MOS defined in ITU-T Recommendation P.800.1 [7], whereas the objective measurements are specified in ITU-T Recommendation P.862 defining "Perceptual evaluation speech quality (PESQ)" [8].

MOS is defined as a statistical average of qualitative ratings of test sentences heard over the phone as perceived by a panel of test listeners. The MOS score ranges from 1 for “bad” to 5 for “excellent” with respect to commonly specified criteria. This method of voice quality assessment is highly subjective, labor intensive and inadequate for frequent voice quality testing. To overcome the deficiencies of subjective MOS scoring and to computerize the said deficiencies, the Perceptual Evaluation Speech Quality (PESQ) algorithm was developed and standardized. The algorithm is based on a measurement of distortions of voice signal passed through a communications system under testing.

There exist many professional tools for metrics-based objective measurements. They can be classified in the following two classes: hardware/software test instruments and software-only test packages. The hardware/software test instruments provide a very efficient way of speech quality assessment. They contain modules for testing traditional phone networks and VoIP telephony, but their high cost makes them unattractive for the use in academic research.

There exist two basic approaches in software-only test packages. In the first approach, the speech quality is extracted from the analogue voice signals, while in the second one the voice packages are extracted directly on IP level, which enables more accurate measurement and analysis of communication parameters like IP timing, etc. An example of analogue software test package is the Opticom Opera Test Suite [10] and for digital speech quality measurements the WireShark open source software package.

B. The PESQ method

The PESQ method evaluates the quality of the speech signal by comparing the reference signal with the degraded signal. The PESQ algorithm models the human perception of the speech signal and thus enables the prediction of speech quality comparable to the subjective assessment as it would be performed by the human audience.

The structure of the PESQ method is presented in Fig. 1. PESQ consists of several signal processing stages: level aligning and filtering, time alignment, auditory transform, cognitive modeling and prediction of speech quality.

First, the reference and the degraded signal are *level aligned* to a standard listening level and filtered to model a standard telephone handset. Next, the signals are aligned in time.

The time alignment techniques are based on the assumption that the delay of the system is constant in time for a given section of signal i.e. piecewise constant. The piecewise constant delay assumption appears to be valid for many applications, including common variable delay communications systems such as VoIP [11]. The time alignment procedure consists of the next steps. Both signals are narrowband filtered in order to emphasize perceptually important parts. Next, the reference signal is divided into utterances of at least 300 ms duration, containing no silent period longer than 200 ms. Silent periods of the reference signal are identified by a voice activity detector with an adaptive threshold to make the speech/non-speech decision robust to noise. After the signal division, a crude delay estimate is calculated across the entire signals using the envelope correlation method. After eliminating this delay, fine delay and confidence estimation is performed by applying the weighted histogram method. Finally, each utterance is divided in two and each section is processed through the same crude/fine delay estimation stages as before. This is repeated at a large number of division points, until there is no evidence for a delay change 4 ms or greater. These give a delay estimate for each utterance, which is used to find the frame-by-frame delay for use in the auditory transform.

The auditory transform is a frame-by-frame representation of perceived loudness in time and frequency on modified Bark scale. A Fast Fourier Transform (FFT) with a Hamming window is used to calculate the instantaneous power spectrum in each frame, for 50% overlapping frames of 32 ms duration. After frequency and gain variation equalization of the reference and the degraded signal the Bark spectrum is mapped to (Sone) loudness, including a frequency-dependent threshold and exponent. This gives the perceived loudness in each time-frequency cell.

Since the time alignment in certain cases may fail to correctly identify a delay change, it can result in large errors for each section with incorrect delay. Each bad section is then *realigned* and the disturbance recalculated.

Disturbance processing and cognitive modeling is the final stage of the PESQ method. The measure of audible error is calculated from the absolute difference between the degraded and the reference signals. A non-linear average result over time and frequency is calculated after omitting the disturbances which are inaudible.

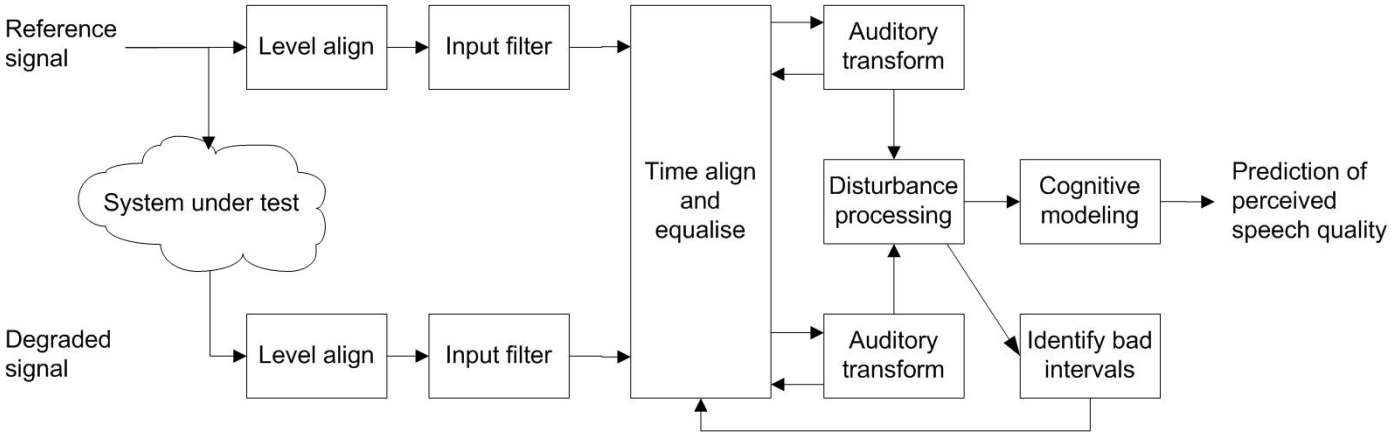

Fig. 1: The structure of the PESQ method [8].

After the aggregation of disturbances in frequency and time, finally, the prediction of perceived speech quality is calculated. The range of the PESQ MOS score is 0.5 for bad to 4.5 for no distortion, although for most cases the output range will be a score between 1.0 and 4.5.

III. Speaker recognition system

A. Speaker recognition tasks

Speaker recognition can be formulated like the process of deciding whether two given speech recordings belong to the same speaker. The use of a machine for speaker recognition is the automatic speaker recognition (ASR) [12]. The ASR is a computing task using speaker voice as personal identifying characteristic. The personal identifying characteristic can be extracted from speaker voice and measured. Measurements in ASR are performed on low-level acoustic features and high-level linguistic features of the speaker. Measurable physical or physiological characteristics are commonly referred as the biometrics. Human beings have many unique personal identity characteristics that make it possible to distinguish one person from another. Many personal identifying characteristics are based on physiological properties, others on behavior, and some combine physiological and behavior properties. Some properties can be perceived very readily such as facial features and behavior. Others, such as fingerprints, iris patterns, and DNA structure are not readily perceived and require biometrics to capture distinguishing characteristics. Speaker's voice is an example of biometric that combines physiological and behavioral characteristics.

The automated process of recognizing a person from his voice includes three different tasks: automatic speaker identification (ASI), automatic speaker verification (ASV) and automatic speaker detection (ASD). The ASI refers to the ability of a machine to uniquely distinguish a person from a larger set of voice samples stored in a voice database without a priori identity claim from that person. On the other hand ASV is the ability of a machine to decide if a speaker is who he claims to be. The third possible task of ASR is the ASD. In the ASD an unknown voice sample is provided and the task is to determine weather or not the one of specified set of known speakers is present in the sample. ASD has been defined in recent years in the NIST speaker recognition evaluations [2]. ASR can be further categorized according to the kind of speech that is input for recognition. If the recognition is performed on the known spoken text at the input and speaker modeling during training has been made for this text, the input mode is text dependent (TD). If, on the contrary, the modeling has been made for unspecified text, the input mode is text independent (TI).

B. Text-dependent vs. text independent speaker recognition

There is important classification between TD and TI speaker recognition [13]. TD speaker recognition utilizes same set of words used during the testing phase and the enrollment phase. In contrast to TD speaker recognition, the TI speaker recognition utilizes any uttered word during enrollment and testing. In other words, TD only models the speaker for a limited set of words in a known context. When the sequence of spoken words is unknown, the problem becomes more difficult and error rates increase.

Error rates of current speaker recognition systems under controlled conditions are low. However, in practical applications many negative factors are encountered including mismatched channel for training and testing, limited training data, unbalanced text, background noise and non-cooperative users [14]. Since the TD speaker recognition systems are mainly used by cooperative users, many of above factors can be avoided. On the other hand, the TI speaker recognition systems are usually subject to non-cooperative users, in fact, sometimes users are even not aware of the fact that they are included in the process of recognizing their identity from their voice, e.g. in forensic applications [15].

The channel variability is one of the most challenging topics in the research of TI speaker recognition techniques [1]. Therefore, for the purpose of this work, we opted for the experiments with employment of TI speaker recognition system.

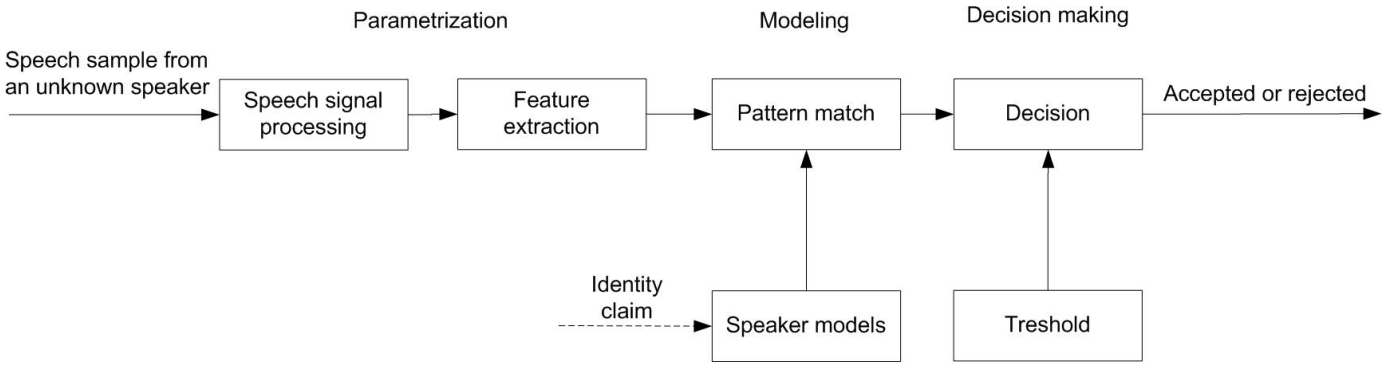

Fig. 2: Blok diagram of a speaker recognition system [4].

C. Speaker recognition process

The process of ASR includes training and recognition phases. The training consists of acquisition of several utterances of known speakers, extracting most representative speaker features, constructing speaker models and storing them in voice pattern database. The training process is also termed as enrolment of the users to the system [16]. A block diagram of a speaker recognition system is shown in Fig. 2. During the recognition phase a sample of speech from the unknown speaker is the input to the system. In case of speaker verification an identity claim is also the input. After the speech sample is recorded and digitized it is prepared for extracting speaker features. Feature extraction is typically some kind of short-term spectral analysis such as filter bank analysis and linear predictive coding (LPC) analysis. During a pattern matching process features of the unknown speaker are compared with the features of known speakers. Finally, the matching score is compared with a predetermined threshold to decide weather two given speech samples belong to the same speaker.

IV. ASRS PERFORMANCE MEASURES

A. Verification and identification

Speaker recognition systems usually comprise verification and identification [16]. Speaker verification is the process of accepting or rejecting the identity claim of a speaker from his voice utterance. In speaker identification, there is no a priori identity claim, and the system determines which speaker provides a given voice utterance from amongst a set of known speakers. In this work the ASRS system performance measurements are based on the speaker verification.

B. Error rates

As any classification system, ASRS also fails in certain number of decisions. There are two types of failed decisions. False acceptance (FA) occurs when the system falsely decides that two speech samples from different speakers belong to the same speaker. As opposite to the FA, false rejection (FR) occurs when the system falsely decides that two speech samples from the same speaker do not belong to the same speaker. ASRS performance is commonly represented as a probability of FA and FR decisions known as false acceptance rate (FAR) and false rejection rate (FRR). Due to practical reasons the use of an equal error rate (EER) as a single number has been established as a good indicator of performance. EER can be found at the operating point where both error rates are equal. However, a single performance number is inadequate to represent the capabilities of an ASRS system in specific applications. Such a system has many operating points, and is best represented by a performance curve [2].

C. DET curves

A tradeoff between FAR and FRR is involved when evaluating the ASRS system. The trade-off between FAR and FRR can be intuitively presented in the form of detection error trade-off (DET) plot [17]. Examples of the DET plot are presented on the Fig. 8 and Fig. 9. In the DET plot we plot error rates on both axes, giving uniform treatment to both types of error. The use of a nonlinear scale for both axes spreads out the plot and better distinguishes different well performing systems and usually produces plots that are close to linear. This scale transforms the error probability by mapping it to its corresponding Gaussian deviate. Thus DET curves are straight lines when the underlying distributions are Gaussian. This makes DET plots more intuitive and visually meaningful.

V. EVALUATION FRAMEWORK

The experimental setup for the evaluation of the the influence of a telephony speech quality on the ASRS performance is presented in Fig. 3. The experimental setup contains two main parts: first, the telephony speech quality test-bed and second, the ASRS with selected speech recordings. The main property of the setup is to enable measurements in two steps. First step is to transmit the selected speech recordings over various telephone networks and measure the speech quality degradations for each of the selected telephone networks under various conditions. Second step is to test the ASRS performance with employment of the degraded speech recordings from the first step. In this section we will describe the evaluation procedure on the speech quality test bed and ASRS with selected speech recordings.

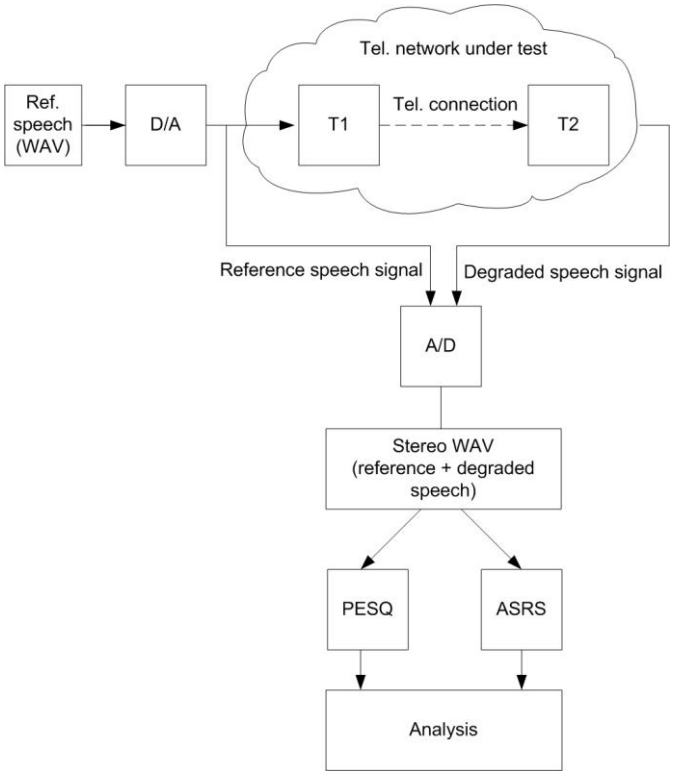

Fig. 3: Evaluation test-bed.

A. Speech quality test- bed

The speech quality test-bed consists of PSTN, GSM and VoWLAN telephony systems and of-line speech quality assessment environment. As opposite to the GSM and PSTN telephony tests, which are performed over live public telephony networks, the VoWLAN setup is built and tested in the laboratory. This enables us to perform VoWLAN testing under different conditions. The speech transmitted over WLAN is degraded by impairments introduced on air and also by the background traffic competing for the same communication medium (for example, IP data terminal and VoIP over WLAN telephone).

Ideally, speech quality testing of the VoWLAN should be carried out with so-called "open-air" measurements at the locations of actual VoWLAN systems. This, however, is not feasible in practice, due to a number of technical and economical reasons such as uncontrollable RF interference and high workload for test execution. Repeatable VoWLAN test results can only be achieved in an environment with tightly controlled RF emission, propagation and reflection. Therefore, the VoWLAN voice-quality testing is usually carried out in RF-shielded chambers with employment of RF signal attenuators, background traffic generator and access points with extra external antennas. Due to limited resources we opted for the simulations of the real-life traffic and limited VoWLAN setups in the open air conditions.

To simulate the real-life traffic and open air conditions we opted for speech quality testing over a range of background bursts in the form of encapsulated RTP traffic and at various distances between wireless access point and clients thus initiating different RF signal attenuation at the tested VoWLAN telephone. The test bed has been partly employed from our previous work [18].

The VoWLAN setup with background RTP traffic is shown in Fig. 4. The single WLAN 802.11b AP is used for the VoIP test connection and the background RTP traffic. The RTP traffic is being transmitted between clients PC#2 and PC#3. For transmitting of the RTP packet streams we used RTP Tools [19]. The automated command line batch procedures controlled by PC#4 initiated the different number of simultaneous RTP streams for each separate test. For the purpose of this work we opted for 4 scenarios, namely 5, 10, 15 and 20 simultaneous RTP streams over the same WLAN channel.

For the speech quality assessments we opted for the PESQ method mainly from two reasons. First, since the PESQ impairment model is very generic and already includes the effects of both packet level impairments (loss, jitter) and signal related impairments such as noise, clipping and distortions caused by coding processes, it is independent from the telephony applications and networks. And second, the PESQ method is standardized and verified in various commercial applications [10].

The speech quality test bed with employment of the PESQ method used in our experimental framework is presented in Fig. 3. The analogue reference voice signal is fed to the telephone handset (T1) and transmitted over the tested telephone network with telephone handset (T2) at the other end of the telephone connection. The degraded voice signal is then digitized together with the reference voice signal at the PC audio card for the off-line PESQ processing, and as we describe in next section, also for ASRS evaluation.

In PESQ processing the analogue reference voice signal from the originating side of the voice connection, represented in standard digital WAV format, is compared to the digitized test voice signal from the other side of this connection and the final PESQ MOS is calculated from this comparison.

Prior to the PESQ MOS calculations the speech recordings from the test data set had to be shortened in order to avoid averaging effect by the PESQ algorithm. Therefore we trimmed each of the recordings in duration of 5 minutes to 5 sections in the duration of 1 minute.

Finally, the analysis of the results and correlations between PESQ MOS and error rates of the ASRS can be observed in the analysis section of the experimental framework.

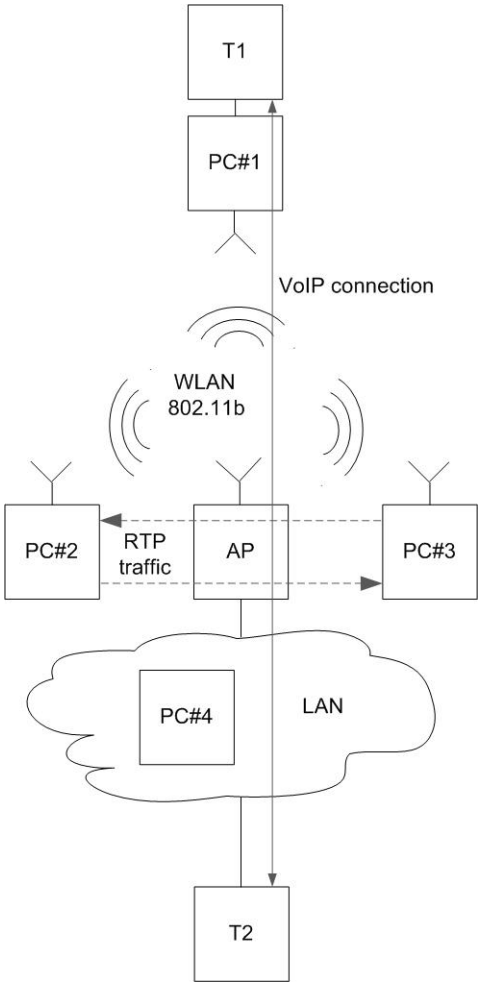

Fig. 4: VoWLAN with encapsulated RTP background traffic.

B. ASRS and selected testing data

The basic platform for evaluating the error rates consists of the ASRS and a dedicated audio corpus of speech recordings. While the ASRS was chosen on the commercial of the shelf market [20], the selected audio corpus was extracted out of the NIST 2008 speech database [2].

The primary purpose of the tested ASRS is the speaker detection on the large number of concurrent telephone calls in text-independent speaker recognition mode. As described in chapter 3, text-independent speaker recognition as oppose to the text-dependent is designed for operation independently of the spoken text, for example ordinary telephone conversation.

NIST 2008 speech database contain large amount of recorded speech in different data sets. Different data sets include various conditions and circumstances for the collected data such are different recording channels (microphone, telephone), different types of speech (conversational speech, interview) different speaker populations (gender, spoken language) and different lengths of recorded samples. Different data sets are usually combined in various tests in order to evaluate systems for different purposes and data conditions.

Typically, each data set selected for the ASRS evaluation contains three separate subsets containing training data, testing data and calibration data. Training and testing data should contain enough audio for training voice signatures and for testing. Additionally, the calibration data should contain enough audio of general speakers not included in test or audio data.

For the purpose of this work we selected 540 English spoken females recorded during conversation over the telephone connection. The training and testing population consists 280 speakers, and the calibration population consists of remaining 260 speakers. The amount of audio for calibration is in duration of 5 minutes of recorded speech for each of the speakers. The training data consist of different amount of data for each speaker. All the selected recordings in the data set are in duration of 5 minutes. The amount of audio for testing is one recording per speaker. The training data consist of different number of recordings for the speakers as follows: 168 speakers with 2 recordings, 103 speakers with 3 recordings, 43 speakers with 4 recordings, 41 speakers with 5 recordings, 3 speakers with 7 recordings, 3 speakers, each with 9, 10 and 28 recordings separately.

C. The ASRS performance evaluation procedure

The ASRS performance evaluation procedure includes preparation of data, the background model creation, enrollment (training voice prints for all client speakers), testing and analysis.

In this work we used data selected as described in previous section. All the recordings from the test data set were previously degraded in the telephony systems as described in section 5.

For the creation of background model we opted for the the GMM algorithm since it has been proven it gives best results for text-independent ASRS [21]. Since background model comprises the features of the target population as they appear in the test data set, it is an important part of an ASRS. Therefore the speech recordings for the background model have to be as much as possible selected out of population with the same spoken language, channel, type of speech etc.

In the testing phase of the ASRS we determined the FAR and FRR of the system for the selected data set. This has been done by comparing the voice-prints created during the enrollment phase to two sets of voice recordings, the authentic (clients) and the non-authentic (impostors). The FRR was determined by observing the system response when comparing the voice prints of the clients to authentic speech recordings. The FAR was determined by observing the system response when comparing the voice prints of the clients to non-authentic recordings (impostors). In our case we combined the impostor tests out of the test data by comparing the voice prints of the clients to all the recordings from test data set of other clients except for their authentic recordings. This gives us more than 60.000 impostor tests and provides enough statistical significance for the resulting error rates.

VI. Experimental results and discussion

In this section we represent the experimental results for (a)

the speech quality assessments of the tested telephony systems and (b) the error rates of ASRS and their correlations by means of DET curves.

A. *Speech quality results*

The PESQ average results with average, minimum and maximum MOS as obtained from several thousand measurements for each of the telephone networks are presented in Table I. As expected, the PSTN outperforms all other telephone networks. For the VoWLAN we observe variations of the MOS from 1.04 to 4.35. As we have shown in our previous work [18], due to the increasing number of background RTP streams, one can observe gradual degradation of average PESQ and at the same time larger spread of PESQ results. The spread of the results for the VoWLAN with excellent signal is clearly visible on the Fig. 5. The variations of the MOS at the lower RF signal for the VoWLAN are presented in the Fig. 6. In the Fig. 7 we observe variations of the MOS for the PSTN which are, as expected, much lower than variations at the VoWLAN.

The variations of the MOS can be attributed to the variations in the speech samples and, for the GSM, slight interference in the local mobile-to-landline interface used in our experimental setup.

Table I: The PESQ results: average, minimum and maximum MOS

	Avg. MOS	Max. MOS	Min. MOS
VoWLAN SE	3.75	4.35	1.04
VoWLAN SL	3.57	4.28	1.11
PSTN	3.95	4.42	2.47
GSM (1 min)	3.18	3.65	2.62

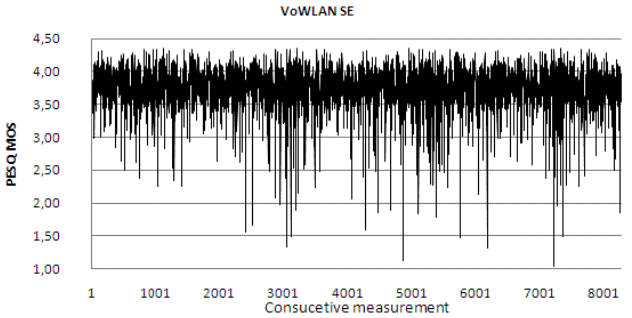

Fig. 5: The PESQ MOS variations for the VoWLAN with excellent signal.

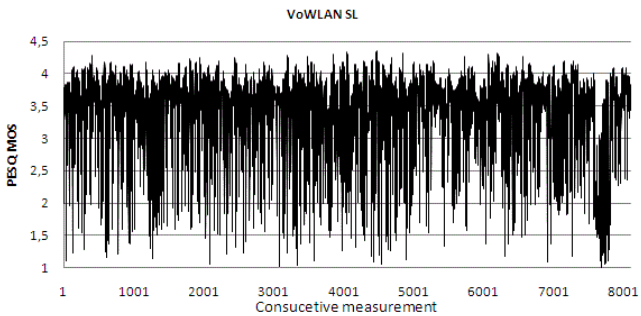

Fig. 6: The PESQ MOS variations for the VoWLAN with lower (-35dB) attenuated signal.

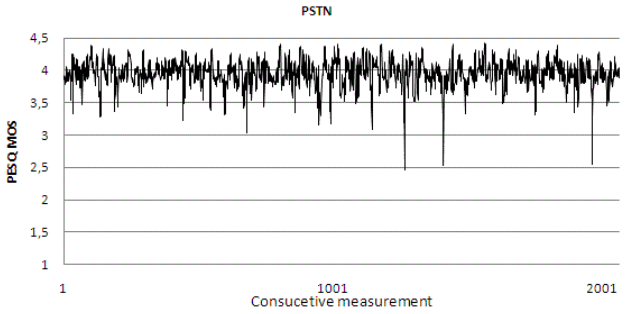

Fig. 7: The PESQ MOS variations for the PSTN.

B. *ASRS error rates and MOS*

Fig. 8 shows the error rates for the evaluation of the ASRS for the VoWLAN. Due to relatively small differences for the error rates with different RTP background traffic we plotted the results from all the tests at 5, 10, 15 and 20 RTP background streams with excellent RF signal (VoWLAN SE) and all the tests with low RF signal (VoWLAN SL) with averaging the results on single plot for VoWLAN SE and VoWLAN SL separately as presented on the Fig. 9.

Fig. 9 represents the error rates for the VoWLAN, GSM, PSTN and original speech recordings. As expected, the original (undegraded) speech recordings outperform the degraded recordings in all telephone networks with EER approximately at 15%. As opposed to PSTN with EER approx. at 18%, the GSM performs slightly worse with EER around 22%. The VoWLAN performance is on average slightly better than the GSM performance. Additionally we observe the effect of signal attenuation on the WLAN with EER difference around 3% in favor of the VoWLAN SE.

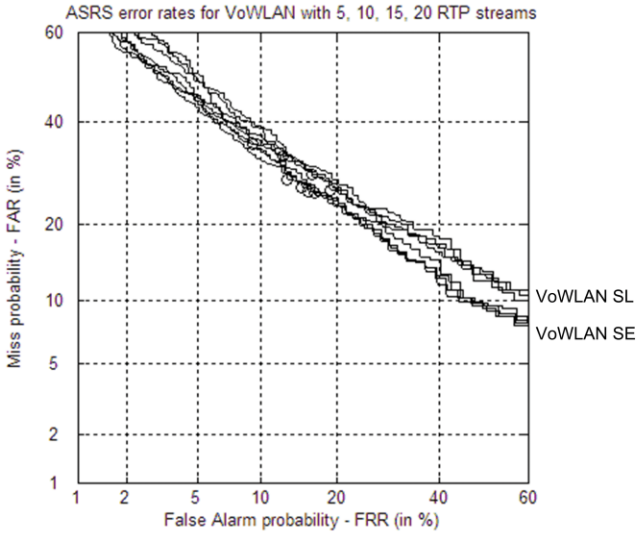

Fig. 8: The error rates of the ASRS system for speech recordings impaired in VoWLAN with different RTP background traffic.

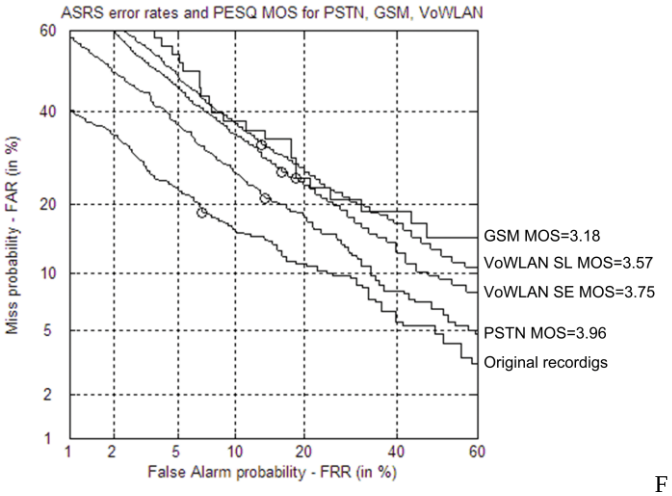

Fig. 9: The error rates of the ASRS system for speech recordings impaired in various telephone networks represented in the form of DET curves.

VII. Conclusions

The influence of the speech quality degradations in the VoWLAN, GSM and PSTN telephony on the ASRS error rates has been investigated. The speech quality degradations were objectively measured using PESQ method and compared to the error rates of the ASRS. Our first results indicate that background traffic with up to 20 simultaneous RTP channels in WLAN on average does not impair the quality of the speech significantly. However, we observed large spread of variations of the MOS. As a consequence the 20 simultaneous RTP background streams do not influence the error rates of the ASRS significantly. However, we demonstrated the ASRS error rates correlate to the speech quality degradations in GSM, PSTN and VoWLAN as measured with PESQ algorithm. The predictions of the expected ASRS error rates with PESQ MOS in the telephony applications could be of great significance. The results show promising approach in order to potentially lower the costs of ASRS evaluations in end user environments. Further work will be oriented towards evaluations with larger data sets under different telephony conditions and employment of analytical tools for data analysis and predictive modeling.

Acknowledgment

The authors acknowledge the contribution of all partners of the WINDECT project and the Forensic Speaker Recognition project.

References

[1] B. Vesničer and F. Mihelič, "The Likelihood Ratio Decision Criterion for Nuisance Attribute Projection in GMM Speaker Verification", *EURASIP Journal on Advances in Signal Processing*, 2008.

[2] D. A. Reynolds, G. R. Doddington, M. A. Przybocki and A. F. Martin, "The NIST speaker recognition evaluation - overview methodology, systems, results, perspective". *Speech Commun*. 31, 2-3 (June 2000), 225-254, 2000.

[3] J. Laver, *Principles of phonetics*, New York: Cambridge University Press, 1994.

[4] J. Benesty, M. M. Sondhi and Y. Huang (Eds.), *Springer Handbook of Speech Processing*, Springer-Verlag, Berlin Heidelberg, 2008.

[5] M. Rainer, U. Heute, C. Antweiler, *Advances in Digital Speech Transmission*, John Wiley & Sons, Ltd., 2008.

[6] L. N. Mangalagiri and S. K. Kopparapu: "Effect of Noise-in-Speech on MFCC Parameters", Recent Advances in Signals and Systems, 2009, pp. 39-43.

[7] ITU-T Recommendation P.800.1, Mean opinion score (MOS) terminology.

[8] Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P., Perceptual Evaluation of Speech Quality (PESQ) - A New Method for Speech Quality Assesment of Telephone Networks and Codecs. In Proceedings of the IEEE International, Conference on Acoustics, Speech and Signal Processing (ICASSP) (Salt Lake City, UT, May 2001), pp. 749–752.

[9] J. Kouril and H. Atassi: "Objective Speech Quality Evaluation. A primarily Experiments on a Various Age and Gender Speakers Corpus", Recent advances in circuits, systems, electronics, control and signal processing, 2009, pp. 333-336.

[10] OPERA PESQ Measurement Software V3.5, [online] Available: http://www.opticom.de.

[11] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual Evaluation of Speech Quality (PESQ): the new ITU standard for end-to-end speech quality assessment, part I—time-delay compensation," J. Audio Eng. Soc., vol. 50, no. 10, pp. 755–764, Oct. 2002.

[12] J. P. Campbell, Jr., "Speaker Recognition: A Tutorial", *V: Proceedings of the IEEE*, vol. 85, no. 9. 1437-1462, 1997.

[13] T. Kinnunen and H. Li, An overview of text-independent speaker recognition: From features to supervectors, Speech Communication 52 (2010) 12–40

[14] R. Saint-nom: "A Shortcut into Speaker Verification", Digest of the proceedings of the WSEAS conferences, 2003.

[15] P. Rose, *Forensic Speaker Identification*, Taylor & Francis, London, 2002.

[16] D. Impedovo and M. Refice: "Optimizing Features Extraction Parameters for Speaker Verification", 12th WSEAS International Conference on SYSTEMS, 2008, pp. 498-503.

[17] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance", *Proceedings EuroSpeech 4*, 1998, pp. 1895-1898. 446, 1998.

[18] R. Blatnik, G. Kandus and T. Javornik, "VoIP/VoWLAN system performance evaluation with low cost experimental test-bed", WSEAS trans. commun., 2007, vol. 6, no. 1, 209-216, 2007.

[19] H. Schulzrinne, P. Pan, A. Tsukamoto, D. Sisalem and S. Casner, RTP tools, [Online]. Available: http://www.cs.columbia.edu/IRT/software/rtptools.

[20] SPID Datasheet, [Online]. Available: http://www.persay.com.

[21] Reynolds, T. Quatieri and R. Dunn, Speaker verification using adapted Gaussian mixture models, *Digital Signal Process*. 10 (2000), pp. 19–41.