INFLUENCE OF SPECIFIC VOIP TRANSMISSION CONDITIONS ON SPEAKER RECOGNITION PROBLEM

P. STARONIEWICZ

Institute of Telecommunications, Teleinformatics and Acoustics
Wroclaw University of Technology
piotr.staroniewicz@pwr.wroc.pl

The paper presents the problem of signal degradation in packet-based voice transmission and its influence on the voice recognition correctness. The Internet is evolving into universal communication network which carries all types of traffic including data, video and voice. Among them the Internet telephony, namely VoIP is going to be an application of a great importance and that is why it is so important to assess how specific conditions and distortions of the Internet transmission (speech coding and most of all packet loss and delay) can influence speaker recognition problem. The Gaussian Mixture Models classification, the feature extraction, the Internet speech transmission standards and the signal degradation methodology applied in the tested system were overviewed. The experiments carried out for two most commonly applied encoders (G.711 and G.723) and three network conditions (poor, average and with no packet loss) revealed a minor significance of the packet loss problem in the tested text-independent system.

Key words: speaker recognition, VoIP

1. Introduction

The Internet is evolving into a universal communication network and it is contemplated that it will carry all types of traffic, including voice, video and data. Among them, telephony, namely VoIP (Voice over IP) is an application of a great importance. The automatic, objective speaker identification and verification problems was partly solved for transmission over traditional PSTN networks (Public Switched Telephone Network). It is also important to assess how specific conditions and distortions of the Internet transmission (like packet delay and loss) can influence the speaker recognition problem. Gaussian Mixture Models (GMMs) are dominant classifiers in nowadays text-independent speaker recognition [2, 4] and is used as a generic probabilistic model for multivariate densities. GMM-based systems have been applied to the annual NIST (National Institute of Standards and Technology) Speaker Recognition Evaluation (SRE), which has produced the state-of-the-art performance [4]. The advantages of using a GMM are that it is computationally inexpensive and based on a well-understood statistical model. What is the most important for text independent tasks is that the GMM is insensitive to temporal aspects of speech, modelling only the underlying distribution of acoustic observation from a speaker [2].

2. Voice transmission over Internet

The voice degradation during the VoIP transmission appears on three levels: acoustics, coding and packet transmission. Selecting a codec is an essential problem for speech transmission. The codec converts analog voice signal to a digitized bit stream at one end of the channel and returns it to its analog state at the other [6].

Table 2.1. Characteristics of speech codecs used in packet networks.

Codec	Type	Bit rate	Frame size	Total delay
G.711	PCM	64 kbps	Depends on packet size	
G.726	ADPCM	32 kbps		
G.729	CS-ACELP	8 kbps	10 ms	25 ms
G.729A	CS-ACELP	8 kbps	10 ms	25 ms
G.723.1	MP-MLQ	6.3/5.3 kbps	30 ms	67.5 ms
GSM.EFR	ACELP	12.2 kbps	20 ms	40 ms

Table 2.1 shows typical voice over IP codecs [6, 9]. The G.711 codec provides a high quality connection with the PCM (pulse code modulation) coding. It is a waveform codec which operates at 64 kbps and which packet size is set arbitrary (for 20ms packetization the delay is 20ms). The G.726 codec is also a waveform codec which also has the packet size set arbitrarily. It reduces the data rate (degrading the quality) and uses the ADPCM (adaptive differential pulse code modulation). For both above codecs the processing delay is negligible and the main delay associated with the use of them is the packetization delay. This is equivalent to the packet length which is usually from 10 to 40 ms.The CELP (code excited linear predictive) codecs are based on the acoustic model of the vocal tract during the speech production which makes the transmission with a lower data rate possible (typically from 4 to 16 for telephony applications). Therefore CELP codecs create more delays than waveform codecs. The G.729 is the 8 kbps codec with good delay characteristics (due to a short frame) and acceptable voice quality. The G.729A has a reduced coding complexity and identical decoding with the equivalent voice quality in comparison to the above. The G.723.1 codec based on multi-pulse maximum likelihood quantization is applied in bandwidth limited transmission channels. The GSM.EFR is a wireless codec which uses a 20 ms frame length. Beside speech coding, the quality of VoIP is determined mainly by packet loss and delay. If a packet is lost the quality degrades and on the other hand, if a packet delay is too high and misses the playout buffer, it leads to a late loss. If a packet is lost or has a large delay, the next one is also likely to do so. The end-to-end packet delay, also known as latency, includes time taken to encode the sound as a digital signal, the signal's journey through the network and the regeneration of it as a sound at the receiving end. Descriptions of the components contributing to the end-to-end delay are presented in Table 2.2. In the IP network packets travel independently and they are interspersed with packets from other network traffic along the way. There are two ways of a packet loss. First, they can be lost at network nodes because of an over-flow in the buffer or because a congested router discards them. Second, packets can be delayed if they take a longer route causing that they can arrive after the prescribed delay and lose their turn.

Table 2.2. Types and causes of packet delays.

Delay sources	Ranges	Description
Transmission Delays	1-100 ms for terrestrial; ~300 ms for geostationary satellite	From short local propagation delays to longest around globe
Sender Delay		
Codec	2-100 ms	Includes encoding and packetization delay, for single IP hop, one frame per packet
Other DSP	0-30 ms	PLC, noise suppression, silence suppression, echo cancellation
Receiver Delays		
Delay for jitter buffer	1-20 ms	depends on utilization and whether congestion control is used
Multiple frames per packet	10-60 ms	Time of additional frames beyond one
Interleaving	5-90 ms	depends on size of frames and packets

Studies on the distribution of the packet loss on the Internet [1, 5, 6] have concluded that this process could be approximated by Markov models. The two states Markov model, also known as the Gilbert model (Fig.2.1) is used most often to capture the temporal loss dependency. In Fig.2.1, p is the probability that the next packet is lost, provided that the previous one has arrived, q is the opposite and 1-q is the conditional loss probability. A more general nth-order Markov chain can also be used for capturing dependencies among events. The next event is assumed to be dependent on the last n events, so it needs 2^n states. Usually it is enough to use up to six states but sometimes it can be 20 to 40. In the Markov model all the past n events can affect the future whereas in the extended Gilbert (the Gilbert model is a special case of the extended Gilbert model when n=2) model only the past n consecutive loss events can do. That is why it does not fully capture the burstiness or clustering between the loss and inter-loss distance metric. ILD (inter-loss distance metric) can be used to prevent it.

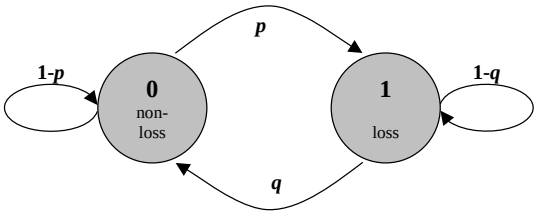

Fig. 2.1. Gilbert model

3. Speaker recognition system

The classical speaker recognition system consists of two main procedures: feature extraction and classification. The MFCC (Mel Frequency Cepstral Coefficients)

parameterization method was chosen [2]. The speech signal is first preemphasized to enhance the high frequencies of the spectrum. After windowing with the Hamming window the signal's fast Fourier transform (FFT) is calculated. Finally the modulus of FFT is extracted and the power spectrum is obtained. To realize the smoothing and get the envelope of the spectrum in an auditory scale (similar to the frequency scale of a human ear) we multiply the spectrum by the Mel scale filterbank. After obtaining the spectral envelope in dB as a final step of parameterization procedure the cosine discrete transform is performed and yields cepstral coefficients. Such received parameters vectors are given to the classification procedure. The GMM [2,4,8] belong to statistical methods of classification. For *D*-dimensional feature vector $\vec{x}$, the mixture likelihood density function is defined as a weighted linear combination of *M* unimodal Gaussian densities $p_i(\vec{x})$:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} w_i p_i(\vec{x}) \cdot \tag{3.1}$$

Each density is parameterized by a *D*x1 mean vector $\vec{\mu}_i$ and *D*x*D* covariance matrix Σ_i:

$$p_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} \mid \Sigma_i \mid^{1/2}} e^{-(1/2)(\vec{x}-\vec{\mu}_i)'\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)} \tag{3.2}$$

The mixture weights w_i satisfy the constraint: $\sum_{i=1}^{M} w_i = 1 \cdot$ (3.3)

Given a collection of training vectors, maximum likelihood model parameters are estimated using the iterative expectation-maximum (EM) algorithm [8]. The EM algorithm iteratively refines the GMM parameters to monotonically increase the likelihood of the estimated model λ. Under the assumption of independence feature vectors, the log-likelihood of model λ for a sequence of feature vectors $X = \{\vec{x}_1, ..., \vec{x}_T\}$ is computed as follows:

$$\log p(X \mid \lambda) = \frac{1}{T} \sum_t \log p(\vec{x}_t \mid \lambda) \cdot \tag{3.4}$$

4. Experiments and discussion

The system was tested with the SV_POL database [5, 7] which consists of speech samples of 22 speakers recorded at 16bit/48kHz in acoustically good conditions (a recording studio, microphone Senheiser MKE66). The speech material included isolated digits and vowels, phonetically rich sentences and strings of digits. For tests the original signals were down-sampled to 8kHz and transmitted via two types of encoders typical for VoIP transmission: G.711 with a-law (64 kbit/sec.) and G.723 (5.3 kbit/sec.). The process of packet loss was simulated with the two states Gilbert model (Fig.2.1), where state "0" represents the case when the packet is lost and state "1" when the packet is correctly transmitted. Probabilities *p* and *q* represent going from state "0" to "1" and from "1" to "0". Two conditions were simulated: bad network conditions (p=0.25, q=0.4) and average network conditions (p=0.1, q=0.7) [1, 7]. The packet length was 30ms in both cases. In the front-end procedures of the voice recognition system experimentally selected feature extraction settings were used: pre-emphasis parameter 0.95, window length of 256 samples, overlap of 128 samples and finally the feature vector consisted of 12 MFCC parameters extracted with the bank of 26

mel-filters. The GMM classifier had 16 Gaussian densities. The number of iterations in the EM algorithm was set experimentally for 15. Table 4.1 presents speaker identification scores for two tested speech items ("S"-phonetically rich sentences and "C"-digit string, i.e. credit card number) and three network conditions with no packet loss, average and poor network conditions as defined above.

Table 4.1. Speaker identification scores for G.711 and G.723 encoders for three network conditions.

Encoder	**No packet loss [%]**		**Average network conditions [%]**		**Poor network conditions [%]**	
	"S"	"C"	"S"	"C"	"S"	"C"
G.711	97.18	98.30	97.02	97.72	94.93	96.81
G.723	96.03	92.64	95.52	87.40	99.34	85.30

For both tested coding types (G.711 and G.723) packet loss does not affect the identification scores. For the low bit rate encoder G.723 (5.3kbit/sec.) there is the maximum fall of 11.51%. The scores of G.723 encoder are on average 4% lower than for G.711.

During the second experiment the tests were carried out on a simple speaker verification system with a fixed decision-making threshold. Table 4.2 presents speaker verification scores for three speech items ("C"-digit string, i.e. credit cared number, "S"-phonetically rich sentence, "D"-spontaneous utterance, i.e. date of speakers birth), two encoders (G.711 and G.723) and three network conditions like in speaker identification tests (no loss, average and poor network conditions). As would be expected, the verification scores of the speech item "D" which is a short spontaneous utterance were the lowest. Similarly as in identification tests, the packet loss does not decrease verification scores significantly.

Table 4.2 Speaker verification scores for G.711 and G.723 encoders for three network conditions

Encoder	**Score**	**No packet loss [%]**			**Average network conditions [%]**			**Poor network conditions [%]**		
		"C"	"S"	"D"	"C"	"S"	"D"	"C"	"S"	"D"
G.711	PAR	98.46	83.94	72.50	95.38	72.60	65.27	100	89.07	69.11
	FRR	1.54	16.05	27.50	4.61	27.39	34.72	0	10.92	30.89
	FAR	3.51	3.94	6.03	4.22	5.25	5.46	13.03	16.24	12.49
G.723	PAR	88.86	87.00	69.72	77.99	66.17	62.70	92.79	80.48	75.65
	FRR	11.13	13.00	30.27	22.01	33.83	37.30	7.21	19.52	24.35
	FAR	4.49	4.70	8.39	6.22	10.30	10.96	18.27	25.49	24.04

PAR-proper acceptance rate, FRR-false rejection rate, FAR-false acceptance rate

The increase of proper acceptance rate for poor network conditions in comparison to the average ones is due to fixed verification thresholds and shifting of conditional probability densities, which is most noticeable for G.723 encoder. However, the deduction of both

network condition rates (the proper acceptance rate as well as the false acceptance rate) give similar results for all the tested utterances, which confirms the minority of the packet-loss problem.

5. Conclusions

The results obtained with the tested text-independent system have shown a minor influence of the packet loss problem on both the speaker identification and verification scores (this confirms the results of the authors earlier preliminary identification experiments presented in [7]). The speaker verification tests of VoIP transmission were only partly solved in the presented paper because only fixed-threshold based recognizer with no usage of UBM (Universal Background Model [2]) was performed. Beside expanding the research to other aspects of speech recognition such as speaker verification and authentication, the main topic of further experiments would probably be testing the influence of the packet loss on the text-dependent speaker recognition. Despite the fact that the packet loss problem does not affect the text-independent speaker recognition scores, it has probably a bigger impact on the text-dependent recognition, which is similar to the automatic speech recognition, more sensitive to time distortions (including packet loss) in a speech signal.

References

[1] BESACIER L., MAYORGA P., BONASTRE J.F., FREDOUILLE C. *Methodology for Evaluating Speaker Robustness over IP Networks*, Proc. of 1st COST 275 workshop in Rome, Italy, 2002, p.43-46

[2] BIMBOT F., BONASTRE J. F., FREDOUILLE C., GRAVIER G., MAGRIN-CHAGNOLLEAU I., MEIGNIER S., MERLIN T., ORTEGA-GARCIA J., PETROVSKA-DELACRETAZ D., REYNOLDS D. A., *A Tutorial on Text-Independent Speaker Verification*, EURASIP Journal on Applied Signal Processing 4, 2004, p.430-451.

[3] EVANS N., MASON J., AUCKENTHALER R., STAMPER R. *Assessment of Speaker Verification Degradation due to Packet Loss in Context of Wireless Devices*, Proc. of 1st COST 275 workshop in Rome, Italy, 2002, p.47-50

[4] REYNOLDS D. A., QUATIERI T. F., DUNN R. B., *Speaker Verification Using Adapted Gaussian Mixture Models*, Digital Signal Processing 10, 2000, p.19-41.

[5] STARONIEWICZ P., *Creation of Real Conditions VoIP Database for Speaker Recognition Purposes*, Proc. of 2nd Cost275 Workshop in Vigo, Spain, 2004, p.23-26.

[6] STARONIEWICZ P., MAJEWSKI W., *Methodology of Speaker Recognition Tests in Semi-Real VoIP Conditions*, Proc. of 3rd Cost275 Workshop in Hertfordshire, UK, 2005, p.33-36.

[7] STARONIEWICZ P., *Speaker Recognition for VoIP Transmission Using Gaussian Mixture Models*, Proc. of 4th International Conference on Computer Recognition Systems CORES 2005, p.738-745.

[8] VLASSIS N., LIKAS A., *A Greedy EM Algorithm for Gaussian Mixture Learning*, Neural Processing Letters 15, 2002.

[9] ITU-T Recommendation H.323, *Packet-based multimedia systems*.