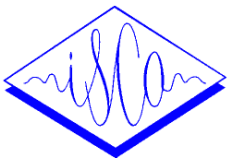

2001: A Speaker Odyssey
The Speaker Recognition Workshop
Crete, Greece
June 18-22, 2001

Speaker Verification over Cellular Networks

Ran Gazit, Yaakov Metzger, Orith Toledo

Persay Ltd.
23 HaBarzel St., Tel Aviv 69710, Israel
Ran_Gazit@persay.com

Abstract

This paper demonstrates the performance gap between speaker verification over land-line telephone networks and speaker verification over cellular networks. The paper shows that the cellular coding accounts for only a fraction of the observed performance gap. A dual-channel corpus, with speakers recorded simultaneously in a land-line phone and a cellular phone, is used to study the effect of the cellular channel on speaker verification performance.

1. Introduction

Classical speaker recognition algorithms were derived using high-quality speech in some cases, and data recorded through land-line telephone networks in other cases. All the standard speech corpora for development and evaluation of speaker recognition systems [1] were recorded through land-line networks, and the yearly evaluations conducted by NIST [2] are based on the Switchboard I-II corpora, recorded through the fixed US telephone network. However, in a couple of years the number of cellular users is expected to be higher than the number of land-line telephone users. By the end of 2005, there will be 326 million cellular users in Western Europe, representing a penetration rate of 83 percent [3]. Can the classical speaker recognition algorithms derived over land-line data work equally well on cellular data?

This paper shows that classical algorithms, which provide excellent performance over land-line data, perform poorly when tested over cellular data of the same speakers. Several previous studies [4-8] attempted to analyze the effect of the cellular channel by transcoding (coding/encoding) land-line data through various types of cellular coding standards, such as GSM and G.723.1. This paper shows that the distortion caused by the cellular coding accounts for only a fraction of the performance gap.

The paper examines some of the causes for this performance gap, using a corpus of speech recorded simultaneously through a land-line phone and a cellular phone. Such a corpus can be used to derive optimal mappings between temporal trajectories of logarithmic energies from both channels, or to derive optimal spectral basis vectors that minimize the spectral distortion caused by the cellular channel. Similar approaches were used in [9,10] for speech recognition, where the dual-channel effect was artificially created by playing high-quality data through cellular phones. In [11] these methods were used for solving the mis-match problem in speaker recognition between carbon-button and electret handsets. However, our study shows that the various types of distortions introduced by the cellular channel diminish the possible benefit of this approach.

The paper is organized as follows: Section 2 describes results of some performance studies on a baseline speaker verification system. The system is tested on land-line data, cellular data and transcoded data (land-line data passed through cellular encoding/decoding algorithms). Section 3 describes the dual-channel corpus and shows some examples of the various distortions, using data recorded simultaneously in land-line phone and cellular phone. Section 4 discusses the cellular channel effect and Section 5 concludes the paper.

2. Baseline performance analysis

In order to estimate the performance gap between speaker verification over land-line telephone networks and cellular networks, we use a multi-channel, text-independent speech corpus, collected by Persay during 1999. The target group contains 43 female speakers and 42 male speakers, each with at least 10 one-minute sessions of free speech recorded through land-line phone, and at least 10 one-minute sessions of free speech recorded through various cellular networks and handsets. The time interval between sessions was several days up to several weeks. All sessions were recorded at 8 kHz through the same telephony board and processed with the same algorithm. Sessions with excessive amount of background noise or intelligible speech were marked by a human listener and excluded from the experiments described below.

The first 3 sessions of each speaker, in each channel separately, were used to train a 30-mixture GMM with a diagonal covariance matrix [12]. The remaining sessions were kept for testing. In addition, four speaker-independent background models were trained, for the cellular/males, land-line/males, cellular/females and land-line/females speaker groups. Each background model is based on more than 60 one-minute sessions of different speakers, not represented in the target groups. Two additional background models combine males data from both channels and females data from both channels. The mixed-channel background models are based on utterances from more than 120 speakers. Each background model is also a 30-mixture GMM with a diagonal covariance matrix.

The features used in this experiment are 12 LPC-cepstrum coefficients appended with 12 delta-cepstrum coefficients, computed over 25ms frames with a frame interval of 12.5ms. An energy-based speech detector was used to discard low energy frames. Cepstral mean subtraction was used to

compensate for stationary channel effects. This specific feature set and model type is not necessarily the optimal configuration in terms of performance in either one of the channels. It is used here as a baseline system only to demonstrate the effect of the cellular channel.

2.1. Fixed channel results

Each speaker is tested with his or her own test sessions (sessions not used for training), and with three test sessions from each one of the other (same gender) speakers. We assume at this stage fixed-channel and fixed-gender environments: cellular/males sessions are tested against cellular/males target models and normalized by the cellular/males background model.

In order to estimate the effect of the cellular coding on speaker verification performance, we have decoded and encoded all the data recorded in land-line phone, through GSM and G.723.1 coding algorithms. The GSM codec is the ETSI GSM 06.10 RPE-LTP full-rate speech transcoder at 12.2 kb/s [13]. The G.723.1 codec is the floating point CELP-based ITU-T standard codec at 5.3 kb/s.

Table 1 shows the equal-error-rate obtained for each channel and gender combination, when test segment length is limited to 30 seconds. The large performance gap between land-line phones and cellular phones is evident. The GSM effect is marginal, as observed in [4], and the performance slightly degrades with the coder bit rate, as described in [5]. However, it is clear that the cellular coding accounts for only a small part of the performance gap between land-line and cellular.

Channel	Males	Females
Land-line phone	2.69	3.15
GSM (12.2 kb/s)	3.14	3.81
G.723.1 (5.3 kb/s)	3.58	4.47
Cellular phone	7.85	9.08

Table 1. Equal error rate in fixed channel, fixed gender scenarios, over the same set of speakers in all channels. GSM and G.723.1 are simulated by transcoding the land-line sessions. Audio length for training is 3 minutes, for testing - 30 seconds.

2.2. Mixed-channel results

Since each speaker in the corpus described above is represented in both channels, both random and controlled mixed-channel experiments can be performed. In controlled experiments, cellular sessions are tested against land-line models, and vice-versa. This represents a worst case scenario where the channel during test is not represented in the model. In random experiments, calls can be selected at random from both channels, with the selection probability representing the expected distribution of channels in a specific operational environment.

Mixed channel experiments raise the question of background channel. Assuming that separate background models exist for each channel (and gender), which background model should be used for normalizing the verification scores?

Table 2 shows the equal-error-rate obtained in controlled mixed-channel experiments. For each gender, three separate background models (BM) were used: one matches the channel used during training, another matches the channel used during testing, and another combines both channels. As expected, mixed-channel results are way worse than the fixed channel results described in the previous section. The table shows that in a mixed-channel scenario, it is best to select a background model which matches the channel used during training.

Train	Test	BM	Males	Females
Land	Cell	Land	13.88	16.83
		Cell	18.57	18.15
		Land+Cell	15.05	16.84
Cell	Land	Land	20.49	18.29
		Cell	17.03	15.73
		Land+Cell	18.88	16.10

Table 2. Equal error rate in mis-matched channel, fixed gender scenarios, over the same set of speakers in all channels. The Background Model (BM) can come from speakers using the same channel as in training, the same channel as in testing, or from both channels.

In most operational environments, information about the channel being used by the speaker is unavailable. Results in this case can be represented by the mixed-channel background model. As expected, results with the mixed-channel model are better than results with the wrong-channel (not used in training) model. However, if the mixed-channel background data could be clustered into channel-specific groups, and the group which best matches the training channel of each speaker would be used for normalizing tests against this speaker, some improvement in results could be gained.

Forming background models and methods for selecting the best background model in a mixed-channel scenario where channel information is unavailable, is outside the scope of this paper and will be discussed in our forthcoming paper.

3. Dual-media corpus

In order to capture the true distortions introduced by the cellular channel, without passing data recorded in land-line phone through the cellular network, we have recorded a unique dual-media corpus. Each speaker was asked to speak freely into two handsets at the same time. One was a cellular handset, and the other – a standard telephone handset, connected to the fixed land-line network. Data was time-aligned using auto correlation of the raw signal.

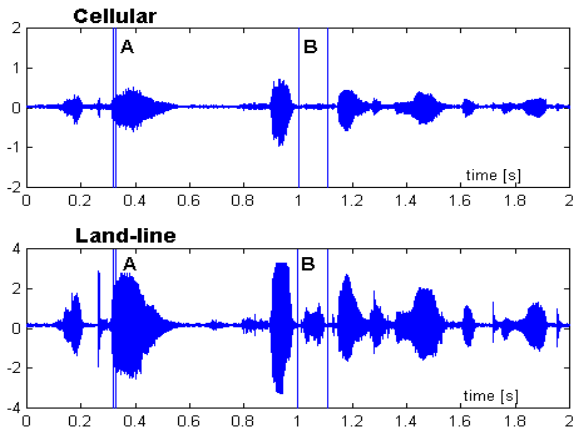

Figure 1: A short phrase, recorded simultaneously through land-line and cellular channels. Area A is voiced – see Figure 2 for details. Area B is unvoiced, and does not pass the cellular channel.

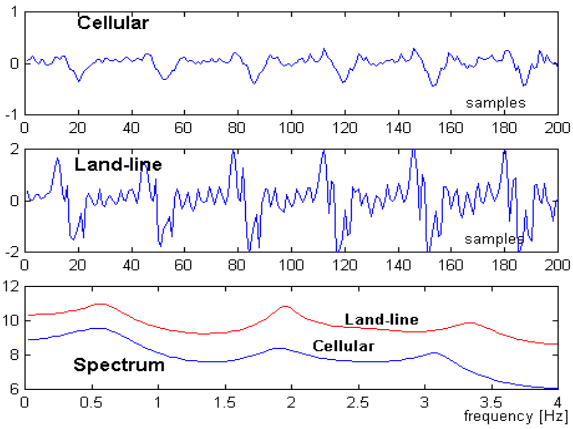

Figure 2: A voiced frame recorded simultaneously through land-line and cellular channels, with the matching LPC- smoothed spectrum. The frame was taken from area A in Figure 1.

The upper part of Figure 1 shows a short utterance in Hebrew, recorded through a cellular phone. The lower part shows the exact same utterance by the same speaker, recorded at the same time (simultaneously) through a land-line phone. Note the area marked as B. This is an unvoiced speech segment, which completely disappeared from the cellular recording. A voiced frame taken from area A is shown in Figure 2, and the LPC-smoothed spectrum of this frame in both channels is shown in the lower part of that figure. The distortion in time and frequency domain is evident from these figures. Note the spectral tilt introduced by the cellular channel, and the apparent shift in the location of the third formant.

Following the approach described by Hermansky et. al. [10,11], the dual-media corpus was used to derive non-causal FIR filters on the time trajectories of logarithmic energies. The filters were designed to minimize the variability introduced by the cellular channel, while preserving the desired signal variability in the fixed-line channel. The undesired variability was estimated from differences between corresponding windows of the spectrum at a specific center frequency, in the cellular channel and the land-line channel. Each window was 1 second long. The desired variability was estimated from differences between far-away spectrum windows in each fixed-line recording.

A similar method can be used to extract spectral basis vectors by looking at the differences between spectral vectors of the same frame of speech that passed through both channels [9]. This method was applied successfully to reduce the effect of mis-match between carbon button and electret handsets.

In either method, the temporal filters or the spectral basis vectors derived from the dual-media database were applied in speaker verification experiments over the corpus described in Section 2. Unfortunately, these approaches could not improve the performance in cellular phone, and had significantly degraded the performance in the fixed-channel land-line case.

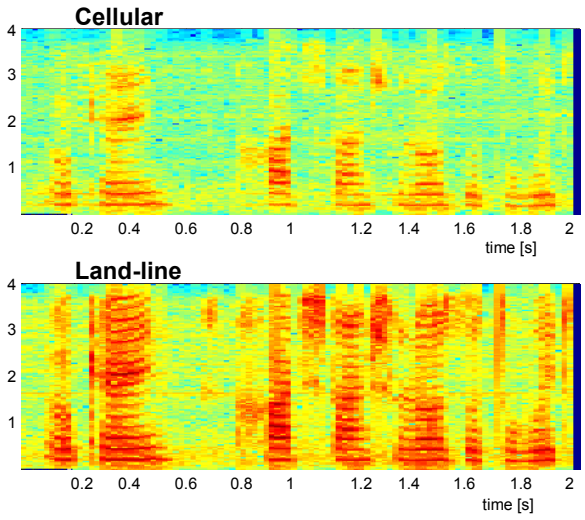

Figure 3: Spectrogram of the phrase shown in Figure 1. Similar energy scales were used for both channels.

4. Analysis of cellular effect

This section outlines some of the possible reasons for the large performance gap between speaker verification performance in land-line and cellular channels. Some of these reasons are due to the different treatment that various sound types get by the cellular channel.

Audio that goes through the cellular network is first segmented into three types of sounds: silence / voiced / unvoiced. Each segment type is treated in a different way by the cellular coder. Silence parts may appear in the decoded signal as perfect silence (DC), or as white noise, sometimes called comfort noise. Standard algorithms for speech enhancement are based

on some form or another of spectral subtraction [14]: the spectrum of the noise / background is estimated from the non-speech parts of the signal, and then subtracted from the speech parts. Such an algorithm will fail over cellular data.

Voiced parts of the signal go through a nonlinear transformation that affects the spectrum shape and moves formants in some cases. Unvoiced parts of the signal are usually coded as white noise driving the LPC model , and may be lost in the decoding process. In general, the higher-frequency contents of the signal are severely affected by the cellular channel. The lack of information in the higher frequency bands of the cellular channel is evident in Figure 3, which shows the spectrogram of the short phrase recorded simultaneously through both channels. This has a severe effect on spectrally-based features, especially with female speakers. This spectral distortion may explain the large degradation in performance, since the higher-frequency bands (>2000Hz) are usually assumed to be more speaker-specific than the lower frequency bands [15].

In addition to the distortions described above, cellular phones are used in many environments, usually outdoors, and the portion of cases with loud background noise is significantly larger than land-line phones, which are being used indoors in most cases. Excessive background noise affects also the speaking style and changes the vocal tract spectrum through the Lombard effect [16]. Other elements that may explain the performance gap are frequent cutoffs, which do not exist in land-line phones, and the miniature, light-weight handset microphones, usually not located in front of the mouth like standard telephone handset.

5. Discussion

This work demonstrated some of the problems associated with speaker verification over celluar networks. These are mainly poor performance in fixed channel environment (using a cellular phone during both training and testing), and unacceptable performance in mixed-channel environments. Another problem is background model selection. This problem is associated with the difficulty in setting a decision threshold that will be channel-independent.

Due to the lack of cellular corpora and benchmark experiments, previous studies had simulated cellular data by passing land-line data through some cellular coding standards, and then checked the performance over the coded/encoded data. This paper shows that the cellular coding can account for only a small fraction of the performance gap.

The results provided in this paper were derived using a very basic speaker verification system. We have studied other combinations of LPC analysis order and number of cepstral coefficients, as well as various other types of features (such as Mel-spectrum and LSF), and could not find a set of features with significant and consistent (in all gender and channel combinations) performance improvement. We have found, though, that some fusion of a cepstrum-based classifier with a prosody-based classifier can provide consistent improvement. This method is currently under study.

6. References

[1] J.P. Campbell and D.A. Reynolds, “Corpora for the Evaluation of Speaker Recognition Systems”, Proc. ICASSP 1999

[2] NIST Speaker Recognition Evaluation Plans: http://www.nist.gov/speech/test.htm

[3] P. Quigley, “Highlights of Western Europe Wireless Market”, WirelessWeek, October 2, 2000

[4] M. Kuitert and L. Boves, “Speaker Verification with GSM Coded Telephone Speech”, Proc. Eurospeech 97, pp. 975-978

[5] T.F. Quatieri et. al., “Speaker and Language Recognition using Speech Codec Parameters”, Proc. Eurospeech 99, vol. 2, pp. 787-790

[6] Febe de Wet, Bert Cranen, Johan de Veth and Louis Boves, “Comparing Acoustic Features for Robust ASR in Fixed and Cellular Network Applications”, Proc. ICASSP 2000, pp. 1415-1418

[7] T.F. Quatieri et. al., “Speaker Recognition using G.729 Speech Codec Parameters”, Proc. ICASSP 2000, vol. 2, pp. 1089-1092

[8] L. Besacier et. al., “GSM Speech Coding and Speaker Recognition”, Proc. ICASSP 2000, vol. 2, pp. 1085-1088

[9] Carlos Avendano and Hynek Hermansky, “On the properties of temporal processing for speech in adverse environments”, Proc. WASPA ’97, Mohonk, NY 1997

[10] Hynek Hermansky, Eric A. Wan and Carlos Avendano, “Speech Ehancement based on temporal processing”, Proc. ICASSP 1995

[11] N. Malayath et. al., “Data Driven Temporal Filters and Alternatives to GMM in Speaker Verification”, DSP journal, vol. 10, No. 1-3, 2000

[12] D.A. Reynolds, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, IEEE Trans. Acoust., Speech and Signal Processing, Vol. 3, No. 1, pp. 72083, January 1995.

[13] Jutta Degener , “Digital Speech Compression: Putting the GSM 06.10 RPE-LTP algorithm to work”, *Dr. Dobb’s Journal*, December 1994

[14] Detlef Hardt and Klaus Fellbaum, “Spectral Subtraction and RASTA-filtering in Text-Dependent HMM-based Speaker Verification”, Proc. ICASSP 1997, vol. 2, pp. 867-870

[15] Laurent Besacier and Jean-Francois Bonastre, “Subband Approach for Automatic Speaker Recognition: Optimal Division of the Frequency Domain”, Lecture Notes in Computer Science (1206), Audio- and Video-based Biometric Person Authentication, Springer LNCS, Bigun et. al., Eds., 1997

[16] Herman J.M. Steeneken and John H.L. Hansen, “Speech Under Stress Conditions: Overview of the Effect on Speech Production and on System Performance”, Proc. ICASSP 1999, vol. 4, pp 2079-2081