The COST-277 speech database

Marcos Faundez-Zanuy (*), Martin Hagmüller (**), Gernot Kubin (**), W. Bastiaan Kleijn (***)

(*) Escola Universitaria Politècnica de Mataró (Spain), (**) Graz University of Technology (Austria), (***) Department of Speech, Music and Hearing (KTH) SWEDEN
faundez@eupmt.es, hagmueller@tugraz.at, g.kubin@ieee.org, bastiaan@speech.kth.se

Abstract. Databases are fundamental for research investigations. This paper presents the speech database generated in the framework of COST-277 "Nonlinear speech processing" European project, as a result of European collaboration. This database lets to address two main problems: the relevance of bandwidth extension, and the usefulness of a watermarking with perceptual shaping at different Watermark to Signal ratios. It will be public available after the end of the COST-277 action, in January 2006.

1 Introduction

Competitive algorithm testing on a database shared by dozens of research laboratories is a milestone for getting significant technological advances. Speaker recognition is one of these fields, where several evaluations have been conducted by NIST [1]. In this paper, we present the COST-277 database, generated by means of European collaboration between three European countries: Spain, Sweden and Austria. However, our purpose is not the collection of a new speech database. Rather than this, we have generated two new databases using a subset of an existing one [2], with the objective to study two new topics that can appear with recent technological advances:

1. The study of the relevance of bandwidth extension for speaker recognition systems.
2. The study of a watermark insertion for enhanced security on biometric systems.

A major advantage of database availability is also to set up the evaluation conditions that can avoid some common mistakes done in system designs [3]:

1. "Testing on the training set": the test scores are obtained using the training data, which is an optimal and unrealistic situation.
2. "Overtraining": The whole database is used too extensively in order to optimize the performance. This can be identified when a given algorithm gives exceptionally good performance on just one particular data set.

Thus, our database includes different material for training and testing in order to avoid the first problem. In addition, the availability of a new database helps to test the algorithms over new stuff and thus to check if the algorithms developed by a given laboratory can generalize their results, even in a new topic framework such as bandwidth extension and watermarked signals.

This paper is organized as follows: section 2 describes the database, and section three provides some experimental results as reference.

2 The COST-277 database

We have generated two new databases using an existing one. Next section describes the original and new databases.

2.1 Original database

Although the original database contains hundreds of speakers, several tasks (isolated digits, sentences, free text, etc.), recording sessions, microphones, etc., we have just picked up a small subset due to the procedure for database generation is time consuming and occupies a considerable amount of data (more than 2 DVD).

We have selected two subsets:

a) ISDN: 43 speakers acquired with a PC connected to an ISDN. Thus, the speech signal is A law encoded at a sampling rate fs=8kHz, 8 bit/sample and the bandwidth is 4kHz.

b) MIC: 49 speakers acquired with a simultaneous stereo recording with two different microphones (AKG C-420 and SONY ECM66B). The speech is in wav format at fs=16kHz, 16 bit/sample, and the bandwidth is 8kHz. We have just used the AKG microphone.

In both cases we have selected the following stuff for training and testing:

1. One minute of read text for training
2. Five different sentences for testing, lasting each sentence about 2-3 seconds.

All the speakers read the same text and sentences, so it is also possible to perform a text-dependent experiment.

2.2 Bandwidth extended database

A speech signal that has passed through the public switched telephony network (PSTN) generally has a limited frequency range between 0.3 and 3.4 kHz. This narrow-band speech signal is perceived as muffled compared to the original wide-band (0 – 8 kHz) speech signal. The bandwidth extension algorithms aim at recovering the lost low- (0 – 0.3 kHz) and/or high- (3.4 – 8 kHz) frequency band given the narrow-band speech signal. There are various techniques used for extending the bandwidth of the narrow-band. For instance, vector quantizers can be used for mapping features (e.g., parameters describing the spectral envelope) of the narrow-band to features describing the low- or high-band [4,5]. The method used in this database is based on statistical modelling between the narrow- and high-band [6].

The bandwidth extension algorithm has been directly applied to the ISDN original database, which is a real situation. However, it is interesting to have a reference of a “real” full band signal (see figure 1 for a conceptual diagram). For this purpose, we have generated a narrow band signal using the full band signal. We have used the *potsband* routine, which can be downloaded in [7]. This function meets the specifications of G.151 for any sampling frequency, and has a gain of –3dB at the passband edges.

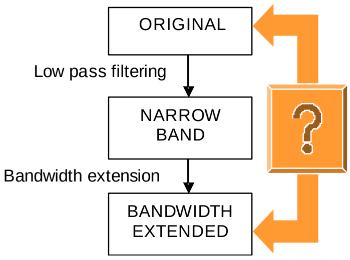

Fig. 1. General pattern recognition system

The bandwidth extension algorithm has been tuned for speech signals with POTS (plain old telephone service) bandwidth, inside the range [300, 3400]. For this reason, we have created the following databases (see table 1):

Table 1. speech databases, fs=sampling frequency (kHz), bps= bits per sample.

Name	Bandwidth[kHz]	fs	bps	description
ISDN	[0, 4]	8	8	Original
ISDNb	[0.3, 3.4]	8	8	ISDN filtered with potsband
ISDNc	[0.1, 8]	8	8	ISDNb + BW extension
MIC	[0, 8]	16	16	Original
MICb	[0.3,3.4]	16	16	MIC filtered with potsband
MICc	[0.1, 8]	16	16	MICb + BW extension

Some experiments with these databases can be found in [8,9].

2.3 Watermarked database

Watermarking is a possibility to include additional information in an audio signal channel without having to sacrifice bandwidth and without the knowledge of the listener. A widely know application of audio watermarking is digital rights management, where the watermark is used to protect copyrights.

Speech watermarking has been used to include additional information in the analog VHF communication channel between pilots and a air traffic controller [10]. Watermarking for biometric signal processing (e.g. speaker verification) can increase the security of the overall system.

Watermarking for speech signals is different than the usual audio watermarking due to the much narrower signal bandwidth. Compared to the 44.1 kHz sampling rate for CD-audio, telephony speech is usually sampled at 8 kHz. Therefore, compared CD-audio watermarking, less information can be embedded in the signal. For perceptual hiding usually the masking levels have to be calculated. The common algorithms used are optimized for CD-audio bandwidth and are computationally very expensive. Another difference is the expected channel noise. For CD-audio the channel noise is usually rather low. Speech on the other side is very often transmitted over noisy channels, in particular true for air traffic control voice communication. On the one hand, the channel noise is a disadvantage; on the other hand this allows much more power for the watermark signal since the channel noise will cover it anyway. The listener expects a certain amount of noise in the signal. A summary of the differences can be seen in table 2. Figure 2 shows an example of a speech frame spectrum with and without watermarking.

Table 2. Audio vs speech watermarking comparison.

	CD-Audio Watermarking	Speech watermarking
Channel noise	Should be very low	Can be high
Bandwidth	Wideband (20 kHz)	Narrowband (4 kHz)
Allowed distortion	Should be not perceivable	low
Processing delay	No issue	Very low (for real time communication)

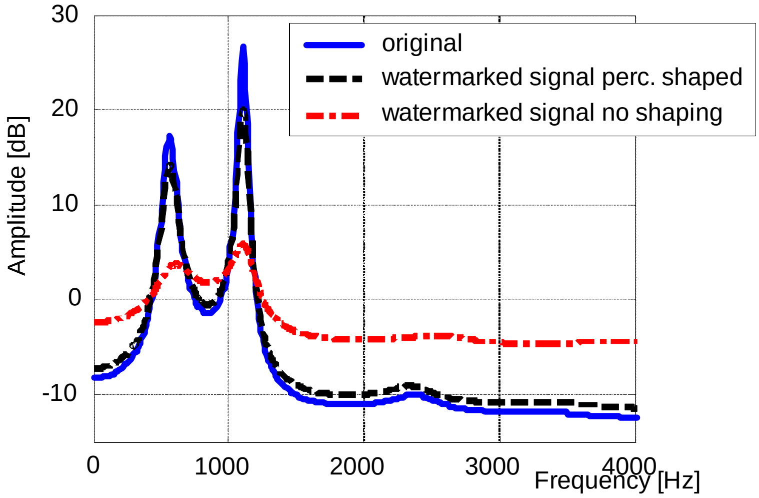

Fig. 2. Example of LPC spectrum envelope a speech fragment, with and without perceptual weighting compared with the original

A more in depth explanation of the watermarking algorithm is beyond the scope of this paper and can be found in [10].

Our previous work [11] stated the convenience for a constant update in security systems in order to keep on being protected. A suitable system for the present time can become obsolete if it is not periodically improved. Usually, the combination of different systems and/ or security mechanisms is the key factor [12] to overcome some of these problems [13-14]. One application of speech watermarking is the combination of speaker recognition biometric system with a watermarking algorithm that will let to check the genuine origin of a given speech signal [15].

Watermark floors higher than the SWR aren't included, since it is not useful.

We have watermarked the MICb database (see table 1) with the following signal to watermark ratios (SWR) and watermark floors (WM floor):

Table 3. Watermark levels (✓ :included, x: not included in the database)

SWR WM floor	0 dB	5 dB	10 dB	15 dB	20 dB
0 dB	✓	x	x	x	x
-5 dB	✓	✓	x	x	x
-10 dB	✓	✓	✓	x	x
-15 dB	✓	✓	✓	✓	x
-20 dB	✓	✓	✓	✓	✓
-25 dB	✓	✓	✓	✓	✓
-30 dB	✓	✓	✓	✓	✓

3 Algorithm evaluation

Speaker recognition [16] can be operated in two ways:

a) Identification: In this approach no identity is claimed from the person. The automatic system must determine who is trying to access.

b) Verification: In this approach the goal of the system is to determine whether the person is who he/she claims to be. This implies that the user must provide an identity and the system just accepts or rejects the users according to a successful or unsuccessful verification. Sometimes this operation mode is named authentication or detection.

In order to evaluate a given algorithm, we propose the following methodology: for each testing signal, a distance measure d_{ijk} is computed, where d_{ijk} is the distance from the *k* realization of an input signal belonging to person *i*, to the model of person *j*.

The data can be structured inside a matrix. This matrix can be drawn as a three dimensional data structure (see figure 3). In our case, *N*=49 and *k*=5.

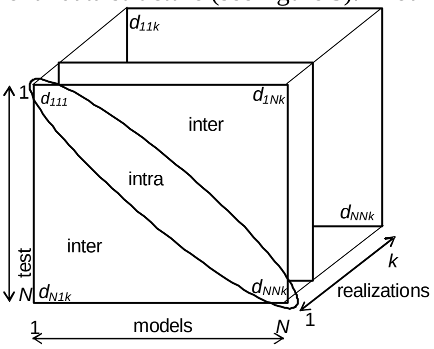

Fig. 3. Proposed data structure.

This proposal has the advantage of an easy comparison and integration of several algorithms by means of data fusion, with a simple matrix addition or more generally a combination. Once upon this matrix is filled up, the evaluation described in next sections should be performed.

3.1 Speaker Identification

The identification rate finds for each realization, in each raw, if the minimum distance is inside the principal diagonal (success) or not (error), and works out the identification rate as the ration between successes and number of trials (successes + errors):

```
for i=1:N,
  for k=1:#trials,
   if($d_{iik}$<$d_{ijk}$) $\forall j ? i$, then success=success+1
   else error=error+1
   end
  end
 end
```

3.2 Speaker verification

Verification systems can be evaluated using the False Acceptance Rate (FAR, those situations where an impostor is accepted) and the False Rejection Rate (FRR, those situations where a speaker is incorrectly rejected), also known in detection theory as False Alarm and Miss, respectively. This framework gives us the possibility of distinguishing between the discriminability of the system and the decision bias. The discriminability is inherent to the classification system used and the discrimination bias is related to the preferences/necessities of the user in relation to the relative importance of each of the two possible mistakes (misses vs. false alarms) that can be done in speaker verification. This trade-off between both errors has to be usually established by adjusting a decision threshold. The performance can be plotted in a ROC (Receiver Operator Characteristic) or in a DET (Detection error trade-off) plot [17]. DET curve gives uniform treatment to both types of error, and uses a scale for both axes, which spreads out the plot and better distinguishes different well performing systems and usually produces plots that are close to linear. DET plot uses a logarithmic scale that expands the extreme parts of the curve, which are the parts that give the most information about the system performance. For this reason the speech community prefers DET instead of ROC plots. Figure 4 shows an example of DET plot, and figure 5 shows a ROC plot.

We can use the minimum value of the Detection Cost Function (DCF) for comparison purposes. This parameter is defined as [17]:

$$DCF = C_{miss} \times P_{miss} \times P_{true} \; + \; C_{fa} \times P_{fa} \times P_{false} \tag{1}$$

Where C_{miss} is the cost of a miss (rejection), C_{fa} is the cost of a false alarm (acceptance), P_{true} is the a priori probability of the target, and $P_{false} = 1 - P_{true}$. $C_{miss} = C_{fa} = 1$.

Nevertheless, this parameter just summarizes the behaviour for a narrow range of operating points in the neighbourhood of the selected threshold. For this reason a whole DET or ROC plot is more interesting for system comparison purposes.

Using the data structure defined in figure 3, we can easily apply the DET curve analysis. We just need to split the distances into two sets: intra-distances (those inside the principal diagonal), and inter-distances (those outside the principal diagonal).

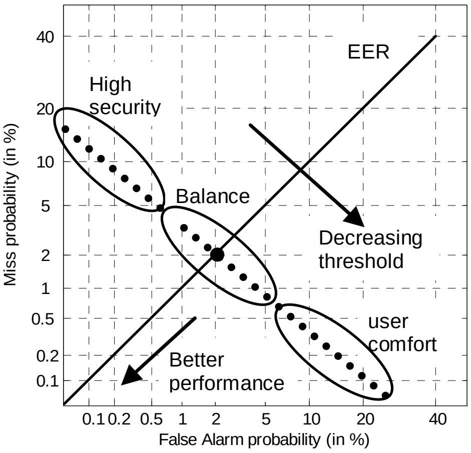

Fig. 4. Example of a DET plot for a speaker verification system (dotted line).

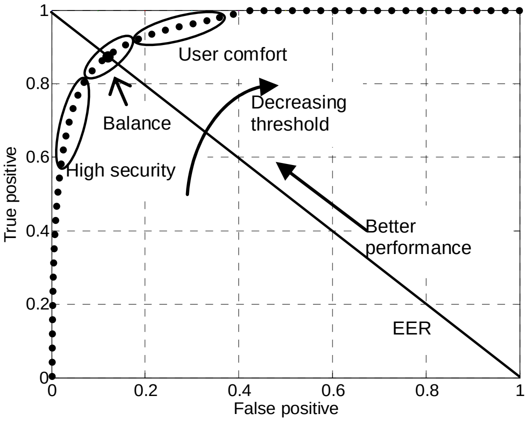

Fig. 5. Example of a ROC plot for a speaker verification system (dotted line).

References

1. http://www.nist.gov
2. Ortega-García J., González-Rodríguez J., and Marrero-Aguiar V., "AHUMADA: A Large Speech Corpus in Spanish for Speaker Characterization and Identification". Speech communication Vol. 31 (2000), pp. 255-264, June 2000
3. Bolle R. M., Ratha N. K., Pankanti S., “Performance evaluation in 1:1 Biometric engines”. Springer Verlag LNCS 3338, pp.27-46 S. Z. Li et al. (Eds.) Sinobiometrics 2004.
4. Enbom N., and Kleijn W. B., “Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients,” in IEEE Workshop on Speech Coding, Porvoo, Finland pp. 1953-1956, 1999

5. Epps J., and Holmes H.W., “A new technique for wideband enhancement of coded narrowband speech,” in IEEE Workshop on Speech Coding, Porvoo, Finland pp. 174-176, 1999
6. Nilsson M., Kleijn W. B., “Avoiding over-estimation in bandwidth extension of telephony speech”, IEEE ICASSP’2001, Salt Lake City, USA
7. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
8. Faundez-Zanuy M., Nilsson M., Kleijn W. B., “On the relevance of bandwidth extension for speaker identification”. Vol. III pp.125-128, EUSIPCO’2002, Toulouse.
9. Faundez-Zanuy M., Nilsson M., Kleijn W. B., “On the relevance of bandwidth extension for speaker verification”. Pp. 2317-2320. ICSLP’2002. Denver
10. Hagmüller M., Hering H., Kröpfl A., and Kubin G, “Speech watermarking for air traffic control”, in Proc. of 12th European Signal Processing Conference, Vienna, Austria, Sept. 6-10, 2004, pp. 1653-1656.
11. Faundez-Zanuy M., “On the vulnerability of biometric security systems”. IEEE Aerospace and Electronic Systems Magazine. Vol.19 nº 6, pp.3-8, June 2004.
12. Faundez-Zanuy M., “Data fusion in biometrics” IEEE Aerospace and Electronic Systems Magazine. IEEE Aerospace and Electronic Systems Magazine, Vol.20 nº 1, pp.34-38, January 2005.
13. Faundez-Zanuy, M., “Privacy issues on biometric systems”. IEEE Aerospace and Electronic Systems Magazine, Vol.20 nº 2, pp.13-15, February 2005.
14. Faundez-Zanuy M., “Biometric recognition: why not massively adopted yet?”. IEEE Aerospace and Electronic Systems Magazine. In press, 2005
15. Faundez-Zanuy M., Hagmüller M., Kubin G. “Speaker identification security improvement by means of speech watermarking”. Submitted to IEEE Trans. On Multimedia.
16. Faundez-Zanuy M., Monte-Moreno E., “state-of-the-art in speaker recognition”. In press IEEE Aerospace and Electronic Systems Magazine. 2005
17. Martin A., Doddington G., Kamm T., Ordowski M., and Przybocki M., “The DET curve in assessment of detection performance”, V. 4, pp.1895-1898, Eurospeech 1997

Acknowledgement

This work has been supported by FEDER and the Spanish grant MCYT TIC2003-08382-C05-02, and the European COST action 277 “Non-linear speech processing”