ACOUSTIC CHARACTERISATION OF SPEECH DATABASES: AN EXAMPLE FOR THE SPEAKER VERIFICATION

M. Falcone and U. Contino
Fondazione Ugo Bordoni, Roma, Italy

ABSTRACT

In this paper we propose a simple set of possible measures to describe speech databases in terms of acoustic features. Features may be 'global' or 'target dependent', i.e. they may or may not be functions of the objective of the corpus design. We focus our attention on the specific problem of speaker verification. In particular we analyse the SIVA database, collected over the telephone line in our Institute during the last summer.

INTRODUCTION

Lasting no more than ten years ago, there were no speech databases available. Although the exigency of speech database was a reality also at that time, only with the success of data driven algorithm in speech research (read HMM and NN), the availability of speech database become a need. The industry also promoted and pushed initiatives in this direction as they well know that no commercial applications are possible without large databases. Thanks to CD technology evolution [1], nowadays there are no difficulties in realising and distributing such databases. The pioneer in this field was the TIMIT. Its prototype was available in 1988, and from that time it is a reference in a widespread research and industry sites.

Today we have more than one hundred of CDs as public database; and many others (probably more than one thousand) have been collected for 'commercial' purpose, i.e. for setting up specific voice applications.

So, in conclusion, we now have a lot of databases. But if we are starting a new research, or we are developing a new application, and we understand that a speech database is needed, do we have enough information to make a choice? Probably no!

In fact the description and the characterisation of a database is usually much more expensive than its 'realisation'. For example you may imagine the effort, in term of man power i.e. in term of money too, needed to make (or just to check) manually a transcription.

Speech technology may overwhelm these problems when word-spotting, automatic text alignment, segmentation, etc. algorithms will reach sufficient performance, higher than the actual one. Today these are far from desired target.

On the other side, a speech database, may be characterised under a pure acoustical point of view. As it is a collection of speech *signals*, these may be characterised objectively, without human decision, simple by well defined measures and algorithms.

Generally speaking we may say that there are three different levels of possible description of a speech database:

- *descriptive*: the design of the collection, the population description, the instrumental set-up, etc.
- *annotation*: all the possible annotations and transcriptions, including word transcription, phonetical labelling, prosodic annotation, etc.
- *acoustical*: the measures related to the physical description of the signal.

Excluding the *annotation*, that is often the most important, expensive and difficult one, it will be commendatory that each database has a detailed *descriptive* and *acoustical* description, in order to make clear its possible use.

We shall explore the "sea of speaker verification", that is a small part of the "ocean of speech technology" in this direction, aiming to define a set of possible measures that should be attached to the speech database, in order to give a clear and useful description of the physical characteristic of the signals.

AVAILABLE DATABASES

As speaker recognition, that includes speaker verification and speaker identification, is just a marginal field, there are few public databases on this topic. Here it is a list of what it is available, i.e. the databases utilised in the most important experiments.

For a more detailed description of these see [2].

TIMIT & NTIMIT

Certainly this is the most famous database. Even if it was designed for speech recognition, it has been widely used also in speaker recognition.

Its telephonic version NTIMIT, has a detailed technical description. This is the unique case of acoustic description, that we know, and it is devoted to describe the transformation of the original database in a telephone quality speech database.

KING

It is the first database designed for speaker verification. It is also famous for the "great-divide", an effect related to some variations in the acquisition instruments. The effect is described in term of system performance, and not in relation to the characteristic of the speech signal (that is of course a more reasonable and interesting description).

YOHO

A database collected under a US federal contract in speaker verification. The public version of this database contains compressed speech file. It is not clear the "degradation" (if any) of the speech after the LPC based compression.

SPIDRE

A selection from the most famous "switchboard" database. Also in this case, there is no acoustical description available.

SIVA

The database we collect over the telephone PSTN line last years [3]. It contains 18 repetitions of 20 male speakers. Each session contains a list of isolated words, a dialogue and a read passage, for a total of about 180 seconds.

It is the one used in this work.

ACOUSTIC CHARACTERISATION: A PROPOSAL

Definition and standardisation of acoustical measures in speech are available only for telephonic speech [4]. Many of these may easily be moved to any other kind of speech signal, but the main problem is: which measures must be performed; using which instruments or algorithms; how the results should be grouped and reported; how to create a 'standard report' that will be easy to use and undertake a familiar look.

This is absolutely not a trivial task, and a definitive and comprehensive definition must be validate by the appropriate international commission and institute as CCITT, NIST, etc.

We do not intend here to give an exhaustive contribution. With this paper we only want to promote this initiative, and give a first contribution in this field. The amount of work and of graphical representation we have done cannot, obviously, be shown here; they will be part of the final release of the SIVA database. It is also our intention to run the same procedures on the previous speech databases, in order to identify different characteristics of the signals.

MEASURES: SOME EXAMPLES

The speech signal we use is a standard 8kHz sampled signal, coded with the American mu-law format. All the analysis are executed on a 256 points window, with a 128 points shift, i.e. with half frame of overlapping. Where spectral transformation is used the signals have been preenphatised with a factor of 0.95, and frames have been windowed using the Hamming mask.

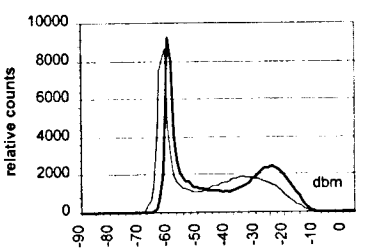

Figure 1. Two speakers' mean energy distribution, speech peaks are at 10dbm

Energy

It is a trivial measure. Nevertheless it is very important that the given values are 'objective', that is no offset is present and the scale reference is correct in relation to the international recommendations. For these reason is quite important that the algorithm respect the CCITT G.711 [5] recommendation, where the numerical values (both in mu-law and a-law) of a

1kHz tone that corresponds to a 0dBm energy are given. Energy normalised histograms clearly give an overview of the recording quality.

These measures should be reported for each session; for each speaker and eventually for 'speaker groups' (e.g. the speakers calling from the same city, or using the same handset, etc.).

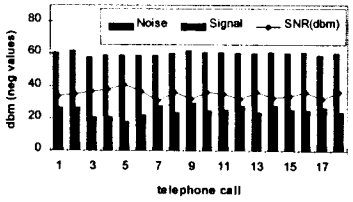

Figure 2. Signal to Noise Ratio (SNR) for one speaker collection

Signal to noise ratio - SNR

It is based on the energy histograms and it is not (unfortunately) an error free measure. More appropriately we can say that it is an estimation, i.e. it is given as the result of the estimation of the mean signal level and the mean noise level. The procedure to measure these mean values range from simple max. estimation to adaptive filtering, and their performance change depending on the speech quality. A human supervision may solve this problem when 'speech' signal to noise ratio is near to the zero value, or when extra signals are added to the speech. Usually, for standard telephone quality signals, automatic methods are adequately. Results may be reported exactly as in the previous case.

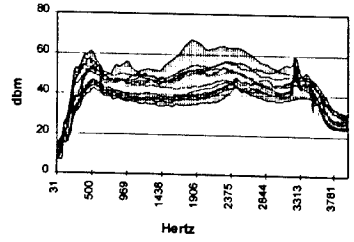

Figure 3. Power spectrum for one speaker, 18 calls existence zone

Long term spectrum

Power spectrum is another classical measure. As for the energy, also in this case it is very important to respect the 'reference signal' so that results may be objectively compared among different databases. This representation is very useful for diagnostic purpose. If the SNR value may insinuate a suspicion, that something is wrong in the signal, the analysis of the long term spectrum will solve in round numbers your doubts. It is difficult to define which kind of 'averages' make sense, as in this specific case a mean over several signals, may mask some important information. So, grouping must be done very carefully and to the averaged spectrum should be added its standard deviation. The first and second order statistical description (mean and standard deviation) of the long term spectra will be, under our experience, sufficient for a diagnostic analysis, if grouping is correct.

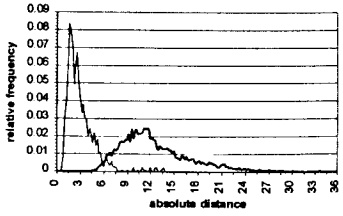

Figure 4. Intra-speaker and inter-speaker variability distribution

Inter-Intra speaker variability

With this measure we are moving towards the specific field of speaker recognition. Inter speaker variability is also important in speaker independent speech recognition, while intra variability is mandatory for speaker dependent speech recognition. In our specific case, they are both crucial. To measure 'variability' a metric, and a matching algorithm must be defined. A plot of an inter or intra speaker variability do not make any sense if the *object*, the *metric* and the *pattern matching* strategy are not defined. Of course comparison between different databases must be done only if these three quantities are the same, otherwise you are not comparing speakers (or signals) but the quality of the speech

model, of the mathematical choices you have done, i.e. your recognition or verification system.

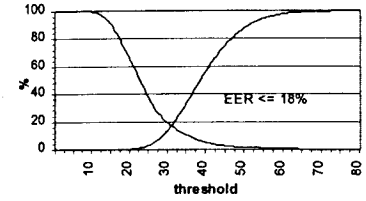

Figure 5. FA and FR using a short utterance (less than 1.5s)

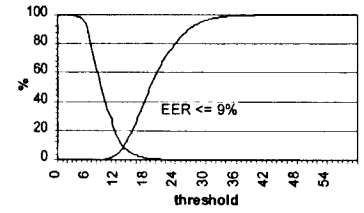

Figure 6. FA and FR using the complete telephone call (about 180s)

"Reference" system performance

It is very important to have an anchor point, a reference that gives the origin of the Cartesian axis where you want to plot the system performance. It seems reasonable that the speaker verification system should be based on the same metrics and parameters used in the inter-intra speaker variability.

Table 1. Performance of the reference system, using 13 LPCC with zero mean

		Training	EER% (FR for FA=0.01)	
		30s	60s	150s
Test	5s	16.5 (47)	12.5 (41)	15.5 (48)
	10s	13.5 (41)	9.0 (25)	13.6 (43)
	15s	8.5 (30)	7.5 (18)	10.7 (32)

As it is always possible to measure inter-intra variability and define a reference system using exactly the same base, it will be foolish to do it in a different way. More difficult is to define reference tests, as usually each database contains a set of files that are not comparable across different databases. The definition of test procedures is probably the most important and difficult point, and we do not address this problem here. In this work we have used a 12th order LPCC parametrisation and a AHSM [6] as distance between two speech samples. We have run several tests; in table one we summarise these reference system performances.

CONCLUSION

Far to define the ultimate recommendations for an acoustical characterisation of speech database, we have outlined the exigency of parting three manifold characterisations of speech databases: one of these is the acoustical description. We suggest a set of possible measure for the speaker verification case, and we report the analysis obtained for the SIVA database. According to the experience we done, these analysis are very useful for the researcher and for the application developer that, starting from the acoustical description, easily obtain a clear and objective view of the characteristic of the database, i.e. check the usefulness of the speech database in relation to his specific purpose.

ACKNOWLEDGEMENT

We would like to thanks all speakers that gave their voices for the SIVA database. They are all friends and colleagues that contribute to the collection without any reward. We are very grateful to all of them for their patience and availability.

REFERENCES

[1] IEEE ASSP Magazine (1985), "*Digital Audio*", Vol.2, N.4

[2] Godfrey J., Graff D., Martin A., Pallet D. (1994), "*Public Databases for Speaker Recognition and Verification*", ESCA Workshop, Martigny, pp.39-42

[3] Contino, U. (1994), "*Una base dati vocale per applicazioni di verifica del parlatore*", (in Italian), FUB Int. Report

[4] CCITT-ITU (1987), "*Handbook on Telephonometry*", Geneve 1987

[5] CCITT Blue Book (1988), "*Pulse Code Modulation of Voice Frequencies*", Rec.G.711, Fascicle III.4, pp.175

[6] Falcone, M., Paoloni, A. (1994), "*Text-Independent Speaker Verification Based on Multiple Reconstruction of Selected Speech Zones*", ESCA Workshop, Martigny, pp.173-176