HYPERBARIC SPEECH UNSCRAMBLING: RESULTS OF AN ANALYSIS/SYNTHESIS METHOD USING PSH/DISPE CDROM SPEECH SAMPLES.

A. Marchal, C. Meunier, D. Masse

Laboratoire "Parole et Langage", URA 261 CNRS
Université de Provence, 29, Avenue Robert Schuman
13621 Aix en Provence, France

ABSTRACT - We describe in this paper the bilingual database of subaquatic and hyperbaric speech (PSH/DISPE), and we present an unscrambling technique which uses an analysis/synthesis method to shift frequencies, to reduce the noise level and to increase speech intelligibility .

INTRODUCTION

Gas mixture, such as heliox, contributed to overcome most of the physiological barriers to effective work in high depth. However this respiratory mixture and increased ambiant pressure modify the spectral characteristics of diver's speech. In addition, the wearing of a facial mask and perturbation of the auditory feeback loop adversely affect the speech production process. As a result, diver's speech is poorly intelligible and communications between divers and surface need to be enhanced. To this end, "voice unscramblers" are being used. However, the technological state of commercially available equipment is dated and the quality of speech remains insufficient.

To help with the design, testing and qualification of new communication devices, a bilingual database (French/English) has been set up and is now available on a CDROM. We describe in this paper both the content and architecture of the resulting PSH/DISPE CDROM and its use as a tool to develop new unscrambling techniques.

1. RECORDINGS

1.1. Speech material

The speech corpus for the database was designed to meet two objectives. First, the database must serve as a tool for fundamental and applied research, and second, it must meet the standards for evaluation of speech intelligibility.

For French, the corpus consists of 4 phonetically balanced lists of 46 words [1] and of 8 sentences where some of these words appear in different syntactical environments. The words have been selected on the basis of their semantic complexity and of their phonological structure. The resulting lists are used for speech intelligibilty tests. The sentences were designed as to allow the study of coarticulatory phenomena.

For English, the corpus consists of the Griffiths lists (4 lists of 50 words, [2]) and of a short passage ("the rainbow...").

1.2. Recordings conditions

The corpus was pronounced twice by each of 17 speakers. They are experienced professional deep divers. They have been selected according to the following criteria: they were 20 to 35 years old; their audiologic control was normal; there was no defect in their speech production and they spoke a standard variety of their native language. The speakers were first trained with the linguistic material by one of our instructor 3 days before the sheduled dive. When they felt comfortable with the word lists and sentences, they were recorded in a quiet room adjacent to the hyperbaric centre. A test was then conducted with one diver to check at ambiant atmosphere the quality of the transmission line from the chamber to the control desk. Three recording conditions were set up:

- in a quiet environment before a dive (i.e. in air)
- during a dive (i.e. under water)
- in the chamber (i.e. during decompression).

In case of error (e.g. forgotten or poorly pronounced words, substitutions), the affected words were repronounced at the end of the recording session.

The recordings using a SONY DAT TCDD10 were done in simulated diving conditions at 60 m, 84 m, 100 m, 150 m, 180 m, 200 m, 250 m, and 300 m (chamber and pool).

Five experimental dives were necessary to collect the data. All the conditions, situations, depths and physical parameters such as gas temperature, oxygen ratio, water temperature, ... were noted.

2. DATA BASE

2.1. Database structure

The database consists of acoustic signal files (40 KHz, 16 bits) of hyperbaric speech and associated files; these later ones provides information about the speakers and the recording conditions.

The PSH/DISPE files are classified according to the *language spoken* (French or English), the *ambient pressure* (atmospheric = 1 bar, or hyperbaric > 6 bars), the *recording location* (chamber or pool) and the dive *depth* .

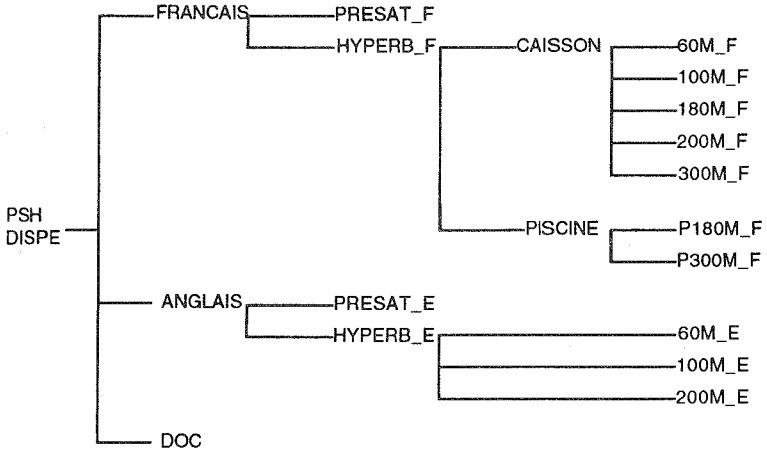

Figure 1: Directory structure of the Database.

Data is put into Database format by re-recording using the EUROPEC software [3]. This operation involves two steps:

First, specification files must be set up containing information about the data in the database. This information concerns the speakers (SPEAKERS.DBF), the corpus (CORPUS.DBF, *.TXT), and the recording conditions (*.RCD).

Then the recordings *per se* are processed.This step involves defining a configuration file for each recording (sampling frequency, type of input, cutoffs for triggering the recording, etc.). The program creates one record (a word or sequence) each time the cutoff defined in the configuration file is reached. Each record in a given signal file is associated with its graphic transcription in a corresponding "orthographic" file.

2.2. Database Management

The file naming system complies with the recommandations of ESPRIT SAM project [4]. The name of the file is composed of 11 characters.

The two first characters (2 letters) refer to the speaker code; the two following (1 letter + 1 digit) refer to the corpus code; and the last four (4 digits), ranging from 0001 to 9999, are the recording numbers.

In the extension, the first character indicates the type of item pronounced, where "S" is for sentences, "P" for paragraphs, and "W" for isolated words. The second character is the language in which the items were pronounced, where "E" is for English and "F" is for French. The third character indicates the type of file, which is either an "S" for signal file or an "O" for an orthographic file. For example, the file "SOM10019.WFS." indicates that: SO is speaking, M1 is the corpus list, 0019 is the recording number; words are pronounced in French and it is a signal file.

To access data in the PSH/DISPE database, the user must have a CDROM drive, a CM153 interface board, and an OROS AU21 or AU22 A/D board. Operations on the database are possible under DOS indicating the path to follow. The access to given files or the search for a specified phonetic context in a precise condition can be highly facilitated by the use of GERSONS database management system [5]. Acoustic analysis is also facilitated by the use of the PTS software package [6], which can be run from within the GERSONS system to directly access given files.

3. SPEECH UNSCRAMBLING: THE ANALYSIS/SYNTHESIS METHOD

3.1. Outline of the analysis/synthesis method

The analysis/synthesis method we used has been first proposed by Mac-Aulay and Quatieri [7]. It supposes. that every frame of the speech signal (which has a finished length L) consists of a finished set of M sinusoïdal components, characterised by their amplitude A_i,their frequency f_i, and their phase f_i. Each frame of analysis can be represented as follows :

$$s(n) = \sum_{i=1}^{M} A_i \cos(2\pi f_i n + \phi_i), \quad n = 0, ..., L - 1 \qquad (1)$$

The interest of this representation is that it gives a better modelisation for speech than for noise, and as a result it increases the Signal/Noise Ratio.

The A_i, f_i, and f_i parameters are extracted from the hyperbaric speech signal thanks to a spectral estimation made with a modified Short Time Fourier Transform. Then, these parameters are transformed from hyperbaric to "normal speech" parameters with which we synthetise the corrected speech signal.

3.2. Parameters estimation

The speech signal is splitted into frames, each of which has a finished length L and each frame is distant from the preceeding one by R samples. Then the frame is multiplied with a low-pass window with appropriate properties [8], and its Discret Fourier Transform is calculated with the Fast Fourier Transform algorithm [9] (the signal is completed to N points with zeros).

We consider the maxima of the modified Short Time Fourier Transform as the sinusoidal components.We simply decided to keep the M highest components after multiplying them with an audition filter :

$$H(f) = 10^{-0.365\, f^{-0.8}} * 10^{-10^{-4} f^4 + 0.65\, e^{-0.6 (f - 3.3)^2}} \qquad f \text{ in kHz} \qquad (2)$$

3.3. The transformation of the sinusoidal components parameters

3.3.1. The frequency transformation

Due to the combined effect of increased pressure and change in the density of the respiratory mixtures, the hyperbaric speech spectrum is modified as expressed by the Fant-Lindquist relation [10]:

$$F_h^2 = K^2 F_a^2 + K^2 \left(\frac{\rho_h}{\rho_a} - 1\right) F_{wa}^2 \qquad (3)$$

where F_h is the hyperbaric frequency, Fa the normal frequency, r_h the helium density, r_a the atmospheric density, K the propagation ratio (He and N), and F_{wa} a constant.

As the spectral estimation gives the hyperbaric frequencies of the sinusoidal components, we have just to reverse the preceeding relation.

3.3.2. The amplitude transformation

We apply in fact two amplitude transformations :

- The vocal tract can be considered as a low pass filter which decreases from 6 dB per octave. Thus, when we transform the "hyperbaric" frequency, we have to transform at the same time its amplitude to take into account this propriety.

- The unvoiced part of the speech decreases more than the voiced part in the hyperbaric environment in a ratio of 10 dB. Thus, we increase the amplitude of the sinusoidal components with this ratio. To do that, we consider that the unvoiced part of the speech corresponds to a part of the signal with little stationnarity. It is detected using a Zero Crossing Detection [11] :

$$\frac{1}{2N} \left(\sum_{i=1}^{N} [\mathrm{sgn}(s(i)) - \mathrm{sgn}(s(i-1))]\right) * Fe \qquad (4)$$

3.4. The synthesis

Because of the discontinuity which can appear between a frame and the following one, we cannot apply directly the relation

$$s(j) = \sum_{i=1}^{M} A_i \cos(2\pi f_i j + \phi_i), \quad j = 0, 1, ..., R - 1 \qquad (5)$$

In fact, we have to smooth the spectrum transitions. This is obtained by binding the sinusoidal components of two contiguous frames using a tracking algorithm [7] which decideswether a sinusoidal component declines, follows or appears between two contiguous frames.

Then we have to interpolate the parameters of the sinusoidal components, and the synthesis is obtained frame by frame with the relation :

$$s^k(j) = \sum_{i=1}^{M} A_i^k(j) \cos\left(2\pi f_i^k(j).j + \phi_i^k(j)\right), \quad j = 0, 1, ..., R - 1 \qquad (6)$$

where k represents the n^r of the frame, i the n^r of the sinusoidal component, j the position in the frame and $P_i^k(j)$ the interpolated parameter.

The amplitude and the frequency are interpolated linearly with the relations :

$$A_i^k(j) = A_i^k + \frac{A_i^{k+1} - A_i^k}{R} j \qquad (7)$$

$$f_i^k(j) = f_i^k + \frac{f_i^{k+1} - f_i^k}{R} j \qquad (8)$$

The continuity between two contiguous frames is made with a constant phase defined by the following recurrent relation :

$$\phi_i^{k+1} = \phi_i^k + 2\pi f_i^{k+1} R \qquad (9)$$

4. EXPERIMENTAL RESULTS

The parameters adopted in our method to unscramble Helium speech can be summarized as follows : We used an analysis window with a length varying between 20 and 40 ms and a Hamming window. The FFT was made on 2048 points.

Although the problem of the compromise between the spectral resolution and the temporal resolution of the FFT can only be optimized and not totaly solved, the results we have obtained are encouraging. As shown on the figure 2 which represents an hyperbaric sentence and the corrected one, the noise level has been greatly reduced by our processing. This is especially true for the high frequencies.

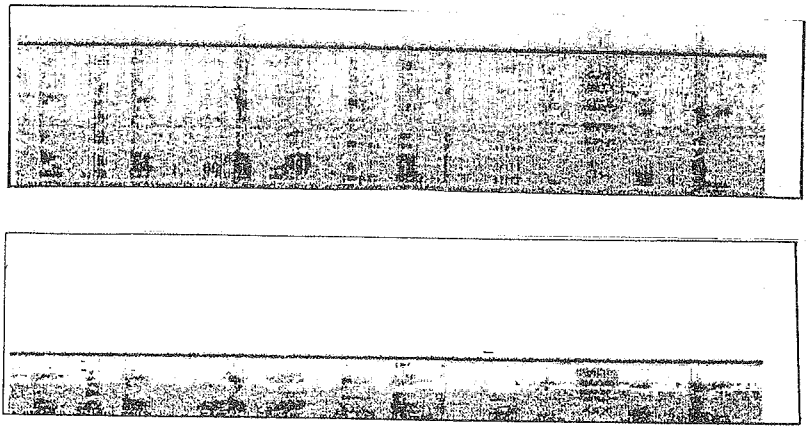

Figure 2: At the top, a sentence pronounced at 300 meters during a dive;
At the bottom, the same sentence after correction

Using the CD-ROM, we have been able to test our algorithm on several speakers in different conditions (speech in chambers, speech during dives, at different depths...)

The enhancement of speech intelligibility appears to be better for speech produced in chambers than during dives, where the additionnal effect of the dead volume of the cavity in the mask creates resonances which are still not corrected by our algorithm.

5. CONCLUDING REMARKS

A speech Database is not an end in itself. It has been used in this work as a tool to help developping a new unscrambling technique.

The improvement of the intelligibility of hyperbaric speech in chambers is good, but it is less convincing during dives.
The method we used is limitated by the necessary compromise temporal resolution - spectral resolution of the FFT. To have a better estimation of the sinusoidal components of the signal, we could possibly use another spectral estimation (as for example the Pisarenko algorithm or the Minimum Cross Entropy Time Frequency Distributions).
It is worth noting that using a DSP and a vectorial processor, the correction processing could be performed in less than two times real time.

Aknowledgments: This research was made possible by grants from MISMER, CNRS, INPP, DRET, CEP&M and Provence-Alpes-Côte-d'Azur Region.

REFERENCES

[1] Marchal, A., Casanova, M.H., Gavarry, P. and Avon, M. "Dispe: a diver's speech data-base", Third Australian International Conference on Speech Science and Technology, Melbourne, 452-457, 1990.

[2] Griffiths, J.D. "Rhyming minimal contrasts: a simplified diagnostic articulation test," *J. Acoust. Soc. Am.* 42, 236-241, 1967

[4] Fourcin, A. J., Harland, G., Barry, W., and Hazan, V. *Speech Input and Output Assessment,* Ellis Horwood, New-York, 1989.

[3] Zeiliger, J., and Sérignat, J.F. "EUROPEC Software (v 4.1), User's Guide Release 4.1," *SAM-ICP-045*, March, Grenoble, 1991.

[5] Foulard, P., Sérignat, J. F. "GERSONS: Système de gestion de la Base de Données des Sons du français. Manuel utilisateur," GRECO-PRC "Communication Homme-Machine" , *Rapport interne,* janvier, I.C.P. Grenoble, 1991.

[6] Caeroux, J. C., and Dolmazon, J. M. "PTS Software V 4.21. User's Manual," *Rapport ESPRIT PROJECT 2589 (SAM)*, june, ICP, Grenoble, 1990.

[7] R.J. Mac Aulay, T.F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoïdal Representation", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-34, pp 744-754, August, 1986.

[8] F.J. Harris, "On the use of windows for harmonic analysis with the Discret Fourier Transform", Proc. of IEEE, vol. 66, N°1, pp. 51-83, January, 1978.

[9] M. Bellanger, Traitement Numérique du Signal, théorie et pratique, Masson, Paris, 1990.

[10] G. Fant, B. Sonnesson, "Speech at High Ambient Air Pressure", STL-QSPR 2, 1964.

[11] R. Boite, M. Kunt, Traitement de la Parole, Complément au Traité d' Electricité, Presses Polytechniques Romandes, Lausanne, 1987.

Note: The PSH/DISPE CDROM is available from CEDROM Technologies, 30 av. de l'observatoire, 75014 Paris; tél.: 33+.16.1.43.35.37.70