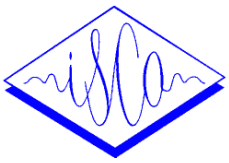
ISCA Archive
http://www.isca-speech.org/archive



2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

TRANSFORMATION OF DATABASES FOR THE EVALUATION OF SPEECH RECOGNIZERS

P. Bardaud, F. Capman, C. Mokbel, C. Tadj and G. Chollet*

Télécom Paris - Dépt. SIGNAL - CNRS URA-820, 46 Rue Barrault 75634 Paris Cedex 13, FRANCE

* IDIAP, Case Postale 609 - 1920 Martigny, CH.

ABSTRACT

This paper proposes two simulation techniques of the Lombard effect. These simulations are based on the Linear Prediction Coding (LPC) and the Linear Multiple Regression (LMR) methods. The LPC model determines a synthesis filter. The LPC coefficients filter are obtained by the processing of the spectral noise signal. This transformation simulates both enhancement spectral tilt and a relative amplification of speech spectral frequency bands where maximum of noise energy exists. This approximates the Lombard effect. These experiments are used to test the limits of SAMREC1 (ENST's DTW recognizer system) in presence of Lombard effect. The second technique is based on the learning of the spectral transformation from the database reference without noise to the same database but recorded with the Lombard effect simulated in the laboratory. This treatment is optimized by dynamic programming.

I. INTRODUCTION

It is useful to compare from an experimental point of view different automatic speech recognizers in order to choose the one which is the most appropriate to a specific application. The best evaluation is obtained by carrying out experiments in real conditions for each recognizer. Unfortunately, this method is not realistic in terms of time, money, and experimental constraints. The use of speech databases aims at overcoming these difficulties, but it remains unrealistic to record speech databases for the evaluation of speech recognizers for all environmental conditions. That is the reason why transformations of speech databases are investigated. This approach is supposed to define what transformation is to be applied to a "reference database" (EUROM0, EUROM1,...) in order to simulate new conditions. Noisy environments are representative of a large number of recognition applications, and it is necessary to have a noise database such as NOISEROM0. Under the hypothesis of additive noise, noisy speech can be obtained by addition of clean speech and noise. But if one wants to be more accurate in simulating noisy speech, the Lombard effect must be taken into account. Speakers modify their speech production in presence of background noise, (figure 1). In this paper, two methods are proposed to simulate the Lombard effect. The first one is based on an LPC noise modeling, while the second one uses a Linear Multiple Regression applied to cepstral vectors.

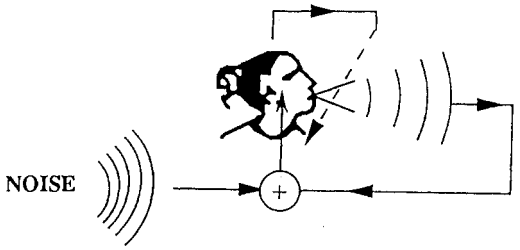


Fig. 1 - noise interaction on the vocal system

II. PRESENTATION OF THE LOMBARD EFFECT

The Lombard effect is now known to have a significant influence on the degradation of speech recognizers performance [1]. In these experiments, Rajasekaran showed that the psychological effect of noise on speech production affects recognition results more than additive noise alone. Some studies have been carried out to clearly define the changes in the speech signal in the presence of background noise. Some interesting results can be found in [2] and [3].

We now give some results of these studies:

1. In noisy environments, speakers increase vocal effort and produce a more energetic speech signal. Amplitude increases but no quantification of this effect has been found.

2. Duration of speech produced in noise tends to increase. This increase is non linear and depends on the phonemes involved.

3. The average fundamental frequency F0 is significantly increased.

4. Upper formants become more intense in the presence of noise, and the spectral slope is enhanced.

5. The main consequence of the Lombard effect on the speech signal in a noisy environment is an improvement of the intelligibility compared "to clean speech" with added noise.

6. In [4], it is shown that the Lombard effect is highly speaker dependent.

10.21437/ICSLP.1992-382

III. LOMBARD EFFECT SIMULATION

We present two techniques that can be used to simulate the Lombard effect, and show how this must be applied to transform speech databases.

The first one is based on an LPC noise modeling and takes into account the enhancement of the spectral tilt.

The second one is a Linear Multiple Regression based transformation applied to cepstral coefficients.

III - 1 Noise LPC modelling:

a) Description:

Experimental results have shown that the signal energy is raised for the frequency band containing noise [5]. In this method, noise is used to introduce a transformation of the spectral slope: it consists of the construction of a synthesis filter from a noise signal.

The different steps are now described:

1) We first computed a noise spectrum from the noise signal using a logarithmic scale.

2) This spectrum is normalized (from 5 dB to 20 dB).

3) A preemphasis is then applied to this normalized spectrum (from 0 to 2 dB), which enables an enhancement of high frequencies and depends on the type of noise.

4) Exponential and inverse FFT are computed to obtain autocorrelation coefficients.

5) The Levinson algorithm is then used to calculate the prediction coefficients Ai of the synthesis filter.

This filter is applied to "clean speech" to give the Lombard effect simulated speech to which noise can be added.

This procedure is presented on figure 2.

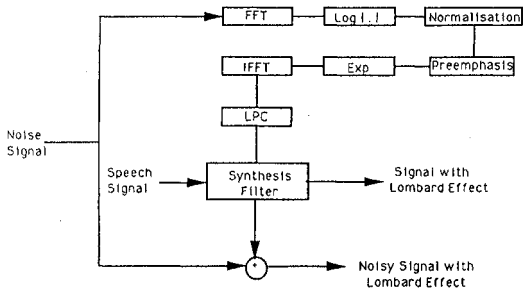


Fig.2 - Lombard simulation by noise modelling

The following curves (figure 3) show the features of a (HF) spectrum noise, normalized and preemphasized spectrum, and synthesis filter result.

b) Evaluation and results of SAM_REC1:

This method has been used to evaluate the SAM_REC1 recognizer [6]. We evaluate the robustness of the system using additive noise with "clean speech" and then with the noisy speech signal with Lombard effect simulation [7].

Experiments were carried out on two speakers from the

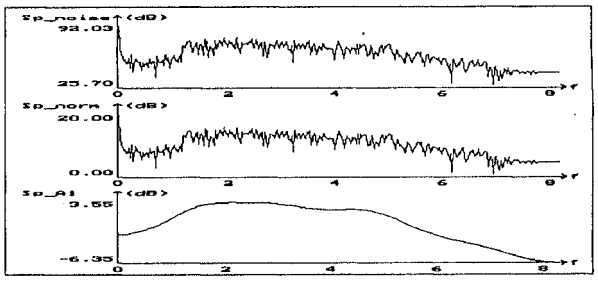


Fig. 3a - preemphasis 0dB

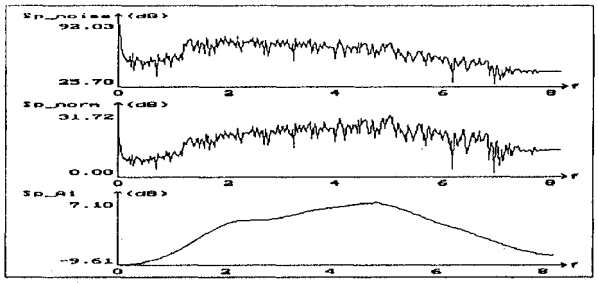


Fig. 3b - preemphasis 2dB

Fig. 3 - Spectral features for HF noise

Top : original noise spectrum
Middle : preemphasis
Bottom : synthesis filter spectrum

EUROM0 database [8] : the male French speaker BG and female French speaker DP. For each speaker, different types of noise were used: white noise, low frequency noise, HF noise, and car noise extracted from one of the European ARS/ESPRIT speech databases. The following curves summarize the results of these experiments with a HF noise signal. The evaluation of the SAM_REC1 recognition rate is shown for different Signal to Noise Ratios (figure 4).

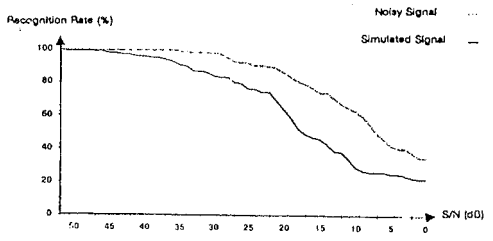


Fig. 4a - HF (1.7KHz) noise with french male from BG Speaker of EUROM0 database

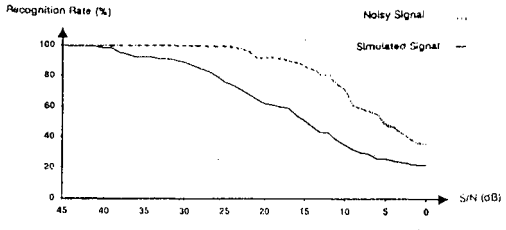


Fig. 4b - HF (1.7KHz) noise with french female from DP Speaker of EUROM0 database

Fig. 4 - Recognition rate curves with the noise modeling Lombard Effect simulation method LPC

The curves show that the recognition rate rapidly decreases. Concerning the evaluation of Lombard simulated speech, improvements have to be made: another choice of the preemphasis value should be considered.

III - 2 LMR based transformation:

a) Description:

We consider here that the Lombard effect can be simulated by a linear transformation of the cepstral coefficients. Before giving more details on the transformation, the analysis / synthesis procedure is presented (figure 5). This is based on an Overlap-and-Add procedure using Short-Time-Fourier-Transform.

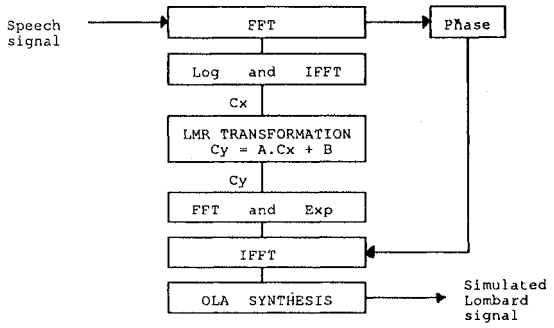


Fig. 5 - Lombard simulation with LMR transformation

1) A Hamming window is applied to each frame of the speech signal and an FFT is calculated. A frame length of 315 samples was used with a shift of 100 samples between successive frames. The sampling frequency was 16 kHz.

2) The logarithm of the amplitude is taken and the phase is stored for later use during synthesis.

3) Cepstral coefficients are calculated, and the transformation is applied.

4) The synthesis stage involves finding the FFT of the cepstral coefficients, taking the exponential of the result and using this in conjunction with the stored phase to resynthesize each frame.

5) The entire signal can then be rebuilt by using Overlap-and-Add on the set of resynthesized frames [9].

The investigated transformation is obtained under the assumption of a linear modeling of the Lombard effect on cepstral vectors. This transformation can be written

$$Cj = A.Ci + B$$

where Ci is the initial cepstrum vector and Cj the transformed one. The matrix A and the vector B define the optimal linear function characterizing the Lombard effect, and are calculated by Linear Multiple Regression.

To evaluate A and B, examples of clean and Lombard speech were recorded for a limited vocabulary uttered by a male French speaker. The vocabulary consisted of the french digits: "zero", "un", "deux", "trois", "quatre", "cinq", "six", "sept", "huit", "neuf", and two control words "debut" and "fin".

To obtain recordings of Lombard speech, the speaker was made to pronounce each word while listening to each of three different types of noise through headphones. Low frequency noise, HF noise and white noise were used. Low frequency noise and HF noise were produced from the white noise using low and high-pass elliptic filters of order 4 with cut-off frequencies of 1.5 kHz and 1.7 kHz respectively.

Cepstral vectors were computed for clean speech and Lombard speech for each type of noise. The resulting cepstral vectors were then aligned using an optimal path obtained by a dynamic programming algorithm applied on MFCC vectors. From these aligned parameters we determined A and B (figure 6).

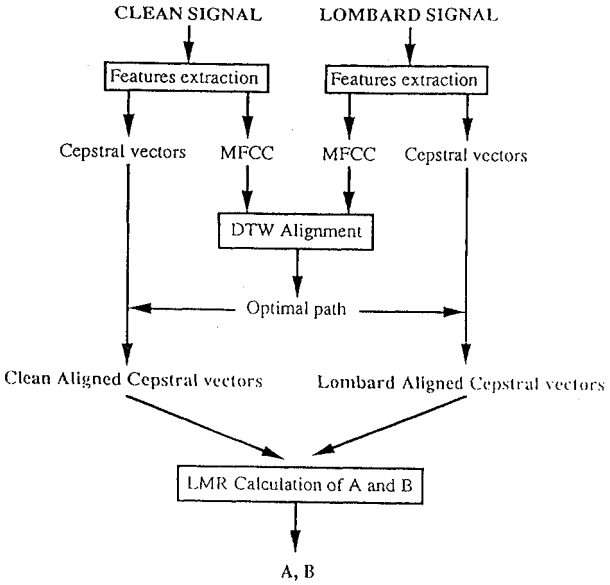


Fig. 6 - Training phase: calculation of A and B using LMR.

In practice, the first 30 cepstral coefficients were used for the evaluation of A and B, but the transformation was not applied to the first cepstrum coefficient, which represents the energy.

b) Evaluation:

This method presents some problems compared to the previous one since, on one hand, examples of Lombard speech are needed for the calculation of the matrix A and the vector B, and on the other hand, the transformation is speaker-dependent.

The calculation of A and B was made using the 10 first words of the vocabulary describe above.

The effect of the transformation can be evaluated by comparing the Euclidean distance between the sets of MFCC vectors (used by SAM_REC1) calculated for each word for clean, Lombard and simulated Lombard speech.

For each types of noise (low frequency, HF and white noise), two sets of Euclidean distances were calculated: the first one between "clean" words and "Lombard" words, and the second one between "Lombard simulated" words and "Lombard" words.

Simulated Lombard words are re_synthesized words from "clean words" using the Linear Multiple Regression transformation.

If we note:

d1_LFN (d1_HFN, and d1_WN respectively) the Euclidean distance between each "clean word" and "Lombard word" for low frequency noise (HF noise, and white noise respectively).

d2_LFN (d2_HFN, and d2_WN respectively) the Euclidean distance between each "simulated Lombard word" and "Lombard word" for low frequency noise (HF noise, and white noise respectively),

the results are on the average:

$$d2\_LFN = 0.82 \cdot d1\_LFN \quad (12 \text{ words})$$

$$d2\_HFN = 0.77 \cdot d1\_HFN \quad (12 \text{ words})$$

$$d2\_WN = 0.74 \cdot d1\_WN \quad (11 \text{ words})$$

Remark: the first word "debut" pronounced by the speaker listening to white noise is more than twice as long as the same word pronounced in quiet environment, that is why no distance can be calculated in this case using our DTW algorithm. This also means that one must take into account in further studies the increase of words duration.

These results have shown that in each case, the distance between Lombard word and Lombard simulated word decreases compared to the distance between Lombard word and clean word. This is particularly true for words used during the training phase. More experiments have to be carried out on more data to confirm these results.

CONCLUSION

This paper describes two techniques to synthesize Lombard simulated speech. These simulations aim at evaluating speech recognizers from a more realistic point of view than only using the hypothesis of additive noise.

The noise LPC modeling method characterizes the energy variations of formants. If the noise presents a flat spectrum, the difference between source speech and simulated one will not be important. Thus, we must determined more precisely the best parameters (preemphasis value and normalization) for each noise speech type.

The LMR transformation method characterizes the frequency variations of formants. However, this transformation needs during the training phase real Lombard data for each type of noise and each speaker.

In order to improve these two methods, and quantify the contribution of each factor to the Lombard effect, a statistical study on real Lombard data has to be carried out. However, the use of a Time-Domain-Pitch-Synchronous-Overlap-and-Add method will enable to simulate non-linear duration modifications and fundamental frequency transformations.

ACKNOWLEDGEMENT

This work was partly supported by EEC under contract ESPRIT-SAM.

BIBLIOGRAPHY

[1] P.K. Rajasekaran, G.R. Doddington and J.W. Picone, "Recognition of speech under stress and in noise", ICASSP, N°14.10, pp. 733-736, 1986, Tokyo.

[2] W.V. Summers, D.B. Pisoni, R.H. Bernacki, R.I. Pedlow and M.A. Stokes, "Effects of noise on speech production: acoustic and perceptual analyses", JASA, Vol. 84, N°3, pp. 917-928, September 1988.

[3] D.B. Pisoni, R.H. Bernacki, H.C. Nusbaum and M. Yuchtman, "Some acoustic-phonetic correlates of speech produced in noise", ICASSP, N°41.10, pp. 1581-1584, 1985.

[4] Y. Anglade and J.C. Junqua, "Acoustic-Phonetic study of Lombard speech in the case of isolated-words", EUSIPCO, 1990.

[5] C. Mokbel, "Reconnaissance de la parole dans le bruit", thèse ENST, Paris, 1992.

[6] G. Chollet, F. Capman and J.F.A. Daoud, "SAMREC1 - Implementation of the ENST-ARECOM reference system on a TMS320C30 - based DSP board", Esprit project 2589 (SAM), final report, March, 1991.

[7] G. Chollet, F. Capman, P. Bardaud and C. Tadj, "Measurements of the limits of the reference recognizer SAM_REC1: noise addition and simulation of the Lombard effect", Esprit project 2589 (SAM), final report, 1992.

[8] Commission of the European Communities, Esprit 90, Project 2589, SAM, "Proceeding of the annual Esprit conference", Brussels, nov. 12-15, 1990.

[9] F.J. CHARPENTIER and M.G. STELLA, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation", IEEE Int-Conf ASSP, Tokyo, pp. 32015-2018, 1986.