EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

NOISY SPEECH ENHANCEMENT BY FUSION OF AUDITORY AND VISUAL INFORMATION: A STUDY OF VOWEL TRANSITIONS

L. Girin, G. Feng & J.L. Schwartz
Institut de la Communication Parlée, UPRESA 5009
INPG/ENSERG/Université Stendhal
B.P. 25, 38040 GRENOBLE CEDEX 09, FRANCE
E-mail : girin@icp.grenet.fr

10.21437/Eurospeech.1997-395

ABSTRACT

This paper deals with a noisy speech enhancement technique based on the fusion of auditory and visual information. We first present the global structure of the system, and then we focus on the tool we used to melt both sources of information. The whole noise reduction system is implemented in the context of vowel transitions corrupted with white noise. A complete evaluation of the system in this context is presented, including distance measures, gaussian classification scores, and a perceptive test. The results are very promising.

1. INTRODUCTION

It has been shown that there exists a complementarity between the auditory and visual modalities of speech [2]. Thus, visual cues can compensate to a certain extent the deficiency of the auditory ones [2][3]. This property is already exploited by bimodal speech recognition systems: their performances are improved by the use of visual data, especially in noisy environments [4].

In a previous paper [1], we presented a new system dedicated to telecommunications or man-machine communication, aiming at enhancing noisy speech with the help of the image of the speaker's face. The purpose was to estimate, from the speaker's lips characteristics, a model of the audio signal, and to filter the noisy audio signal with the estimated model. The results obtained on stationary vowels were very encouraging. But the weak point was that we tried to estimate a complete audio information from the video one, whereas the lip information stays very partial.

In this paper, a new structure for our system is proposed: as an attempt to better exploit the bimodal complementarity, the enhanced auditory information is now estimated from both the noisy auditory and the visual channels. We present first the global structure of the system. Then we focus on the bimodal integration process. Finally, we present some results obtained for the enhancement of vocalic transitions corrupted with additive white noise.

2. STRUCTURE DESIGN

The new system is essentially based on the linear prediction model [5] (fig. 1). First, an LPC analysis is performed on the noisy signal. We obtain spectral parameters and the noisy speech excitation is extracted by filtering through the inverse LPC filter $A_n(z)$. Then, the noisy spectral parameters are combined with the video ones so as to obtain estimated "cleaned" spectral parameters (see section 3). Finally, enhanced speech is synthesised by filtering the excitation through the LPC filter $1/A_e(z)$ derived from the "cleaned" spectral parameters. The whole processing is performed frame-by-frame in the perspective of a continuous speech application.

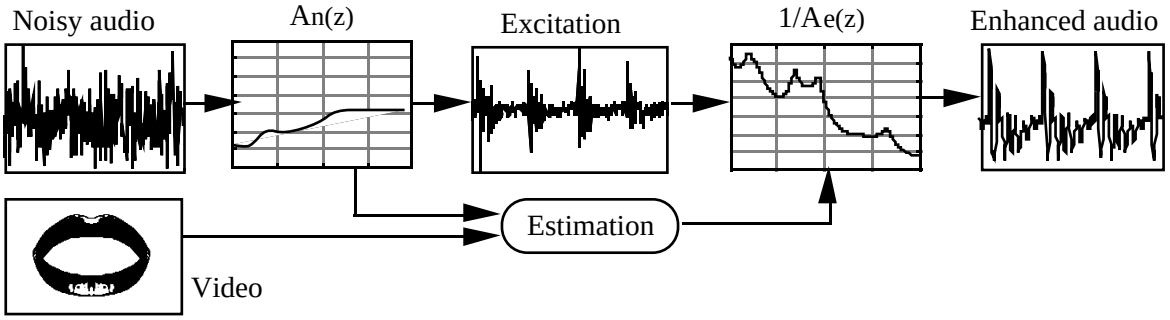

Figure 1 – Structure of the noisy speech enhancement system

3. ESTIMATION OF THE "CLEANED" AUDIO PARAMETERS

The main problem is hence to estimate "cleaned" audio parameters from noisy ones and video ones. We use a linear regression method because of its simplicity and efficiency concerning our problem [1]. The principle is the following. Consider an audio-visual vector as the concatenation of video parameters and audio parameters. Two matrices are built. The first one, called M_{AV}, concatenates the audio-visual vectors issued from a learning corpus where the audio signals are corrupted with noise. The second one, called M_A, concatenates the corresponding vectors of audio coefficients only, issued from the same learning corpus, but in the clean condition. Then we calculate the matrix M that realizes the linear regression between M_{AV} and M_A. Finally, for every new audio-visual noisy vector V_{AV}, the product between V_{AV} and M gives an estimated "cleaned" audio vector V_A.

4. EXPERIMENTATION

4.1. Video and audio inputs

The ICP face processing system [6] allows to automatically extract three basic parameters of the labial contours, namely interolabial width (A), height (B) and area (S). These video parameters are extracted every 20 ms.

The audio information consists in parameters characterising the LPC polynomials. It has been shown that the best performances of the system were obtained with a 50-coefficient spectral representation consisting of the logarithmic values of the 1/A(z) 20-order filter amplitude taken for 50 equally spaced values on the upper-half unit circle. The audio signals are sampled at 16 KHz and the coefficients are calculated on 512 samples (32 ms, which involves an audio window overlap of 12 ms to synchronise with the 20 ms video period).

4.2. The corpus

For stationary vowels, our previous work has given very satisfactory results [1]. In this new implementation, vocalic transitions $V_1V_2V_1$ uttered by one speaker are studied. V_1 and V_2 are within [a, i, y, u]. One item of each of the 16 possible stimuli is used during the learning phase (calculation of the matricial associator), and another one is reserved for the tests described in section 5. With a video acquisition period of 20 ms, we obtain an amount of about 350 audiovisual vectors for a series of 16 stimuli (about 24 frames by stimuli).

4.3. Experimental protocol

We consider only the case of an additive white noise. The results discussed here are obtained with the use of two different matricial associators M: one is dedicaced to enhance stimuli with "strong" SNRs. M is trained with stimuli frames presented at SNRs of ∞, 18, 12, 6 and 0 dB. The other one is dedicaced to enhance stimuli with "small" SNRs. M is trained with stimuli frames presented at SNRs of 6, 0, -6, -12, and -18 dB. During the enhancement process, each frame is submitted to a linear discriminant analysis in order to decide its categorisation in the strong or small noisy condition so that we can choose the corresponding associator. It has been shown that this linear discriminant analysis could separate stimuli with SNR lower than 0 dB or higher than 6 dB with less than 1% errors, while the two associators provide quite similar responses for SNRs between 0 and 6 dB.

4.4. Filtering process

To obtain the filter $1/A_e(z)$ from the "cleaned" spectral parameters, we use an inverse FFT, and apply a 20-order Levinson procedure on the resulting estimated autocorrelation coefficients [5]. During the enhancement phase, both trapezoidal windowing and buffering are applied to the filter junctions to ensure continuity of the enhanced signal.

5. EXPERIMENTAL RESULTS

After an informal qualitative evaluation of the system, three quantitative evaluation procedures are defined: distances measures, gaussian classification test, and perceptive tests. Those evaluations were made with additive white noise for 8 different SNRs (∞, 18, 12, 0, -6, -12, -18 dB).

5.1. Qualitative evaluation

Informal listening tests have revealed a good behaviour of the system. For weak noises, the enhancement does not degrade much the quality of the signals. For medium noises, the effects are more important and useful: the message is easier to understand even if it sometimes sounds surprising (the filtering of the noisy excitation leads to heavy whispered vowels). For highly degrading levels of noise (loss of intelligibility), the system allows to recover the intelligibility of most of the stimuli (almost any [a] or [i], thanks to their distinctive labial shape, with more ambiguity between [u] and [y]).

5.2. Distance measures

The mean Itakura distance [2] has been used to measure the difference between the enhanced and clean spectra. Figure 2 displays the distance calculated on the complete test corpus (16 stimuli) for three conditions: AV stands for the use of audio-visual information, A for audio information only (audio vectors presented at the learning and enhancement phases instead of audio-visual ones), and V for visual information only (visual vectors presented at the learning and enhancement phases instead of audio-visual ones). It has been verified that rather small distances are altogether obtained compared to those between noisy and clean spectra. Hence the procedure does produce a significant enhancement. It can be seen that AV always performs better than A, and almost always better than V (near until -18 dB of SNR). This confirms the interest of visual cues, and the good complementarity between the two modalities.

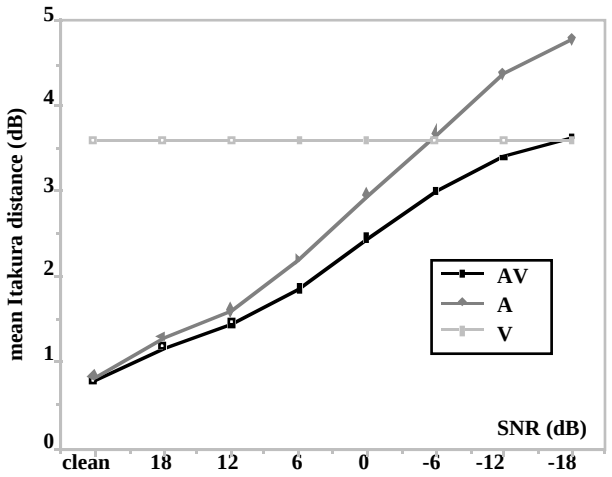

Figure 2 – Mean Itakura distance between enhanced and clean spectra of the test corpus

5.3. Gaussian classification

To evaluate the system in a recognition task, a gaussian classification test has been realised on the four vowels of our stimuli. The items used in this test are two selected frames near each vocalic nuclei of each stimuli. This ensures the absence of coarticulation effects and easy labelling. The video signal is assumed to be quite stable in these selected zones. We obtain 96 items for each level of noise (2 selected frames, 3 vowels per stimuli, 16 stimuli), that is to say 24 per vowel. Since the number of data is small compared to the number of parameters, we reduce the number of audio parameters from 50 to 5 by means of a principal componants analysis (PCA). In the results presented in figure 3, both the PCA and the gaussian classifier parameters are determined with learning data presented at 3 levels of noise (∞, 18, 12 dB). Figure 3 compares the correct classification scores for three conditions: A stands for the use of the noisy audio information only (5 audio parameters) during the learning phase of the classifier. In this condition we have the comparison between A_{noisy} which stands for the use of the noisy test corpus, and $A_{enhanced}$ which stands for the use of the same test corpus after enhancement. In comparison, the AV scores correspond to an audio-visual recogniser applied to a vector combining the video and audio inputs, hence 5 audio and 3 video parameters. Note that all scores are normalised between 0 and 100%, with 0% corresponding to a random choice and 100% to perfect recognition.

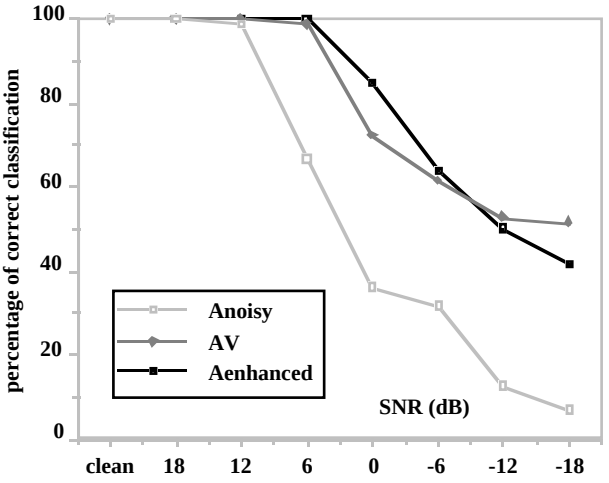

Figure 3 – Gaussian classification test scores

The difference of the scores in the noisy and enhanced conditions confirms the efficiency of the system. Moreover the audio enhanced condition competes well with the audiovisual classification near until SNR = -12 dB.

5.4. Perceptive test

The final evaluation of the system is done with perceptive tests. 17 subjects were asked to identify all stimuli randomly presented in noisy and enhanced condition and for 8 levels of noise. The $V_1V_2V_1$ stimuli were manually segmented into V_1V_2 and V_2V_1 in order to present V_1 only once at each iteration. So, for each point of the identification scores curves in figure 4, e.g. for each level of noise and each condition (noisy or enhanced), we have 1088 responses (16 stimuli with 2 vowels, V_1V_2 and V_2V_1 segments, 17 subjects).

Figure 4 shows that the enhancement is efficient as soon as SNR = 0 dB. The gains obtained (difference between the enhanced and noisy conditions) are about 6% at 6 dB, 17.5% at 0 dB, 18.5% at -6 dB, 30% at -12 dB and reach 42.5% at -18 dB.

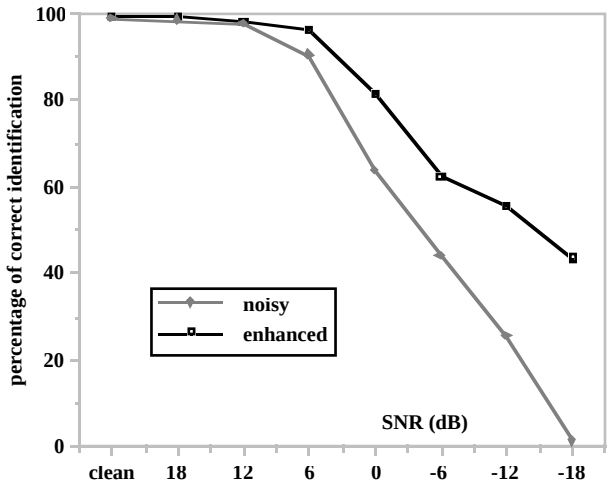

Figure 4 – Perceptive test scores

The confusion matrices for the 5 smallest SNRs are presented in table 1 (the matrices for the 3 stronger SNRs are almost diagonals and present less interest).

		noisy signals				enhanced signals			
SNR	V	a	i	y	u	a	i	y	u
6 dB	a	**271**	1	0	0	**272**	0	0	1
	i	1	**223**	6	2	0	**270**	2	7
	y	0	37	**260**	16	0	2	**260**	11
	u	0	11	6	**254**	0	0	10	**253**
0 dB	a	**271**	2	1	17	**272**	3	0	0
	i	0	**131**	35	17	0	**253**	7	15
	y	1	104	**187**	45	0	14	**222**	70
	u	0	35	49	**203**	0	2	43	**187**
-6 dB	a	**272**	12	17	15	**268**	7	0	1
	i	0	**96**	60	60	4	**243**	28	22
	y	0	128	**155**	89	0	21	**222**	201
	u	0	36	40	**108**	0	1	22	**48**
-12 dB	a	**234**	61	45	42	**247**	6	1	3
	i	27	**108**	94	107	19	**228**	18	25
	y	8	69	**89**	73	6	36	**221**	213
	u	3	34	44	**50**	0	2	32	**31**
-18 dB	a	**128**	129	121	105	**168**	26	4	2
	i	109	**95**	108	113	77	**211**	23	15
	y	24	37	**29**	44	20	33	**218**	225
	u	11	11	14	**10**	7	2	27	**30**

Table 1 – Confusion matrices for the perceptive test. The left matrices are for the noisy condition, the right matrices are for the enhanced condition.

The main performances of our system can be summarized as follows:

1) the disambiguisation of the [i, y] contrast, which is strongly degraded in noise before enhancement. This case represents a good example of the audio/video complementarity of speech (robust video distinction while small audio robustness in noise).

2) the relative disambiguisation of the [a, i] confusion. This appears only for strong noise, since the [a] sound is very robust in noise and the [a, i] lip shapes are not so easily distinctive in dynamic speech compared to static vowels (see [1]). The process works better from [a] to [i] than [i] to [a].

3) the reinforcement of the rounding feature ([y] and [u] are well contrasted with [a] and [i]), which unfortunately leads to the confusion of [u] and [y] for high levels of noise. The poor audio information retrieval in that case remains a weak point of the fusion process.

6. CONCLUSION

We have presented in this paper an original method for noisy speech enhancement using both noisy audio information and the speaker's lip pattern. Its implementation within the scope of vocalic transitions has shown that a good enhancement of the signals can be obtained from the complementarity between the auditory and visual modalities. These results are very promising for the future step of our work, which will involve the dynamic processing of vowel-consonant transitions.

7. AUDIO EXAMPLE

The sound example given contains the three following transitions [aia], [uyu] and [iui]. Each transition is given successively in the noisy and enhanced conditions. The SNR is 0 dB. [sound A0003S01.WAV]

REFERENCES

[1] GIRIN, L. , FENG, G. , & SCHWARTZ, J.-L., Noisy speech enhancement with filters estimated from the speaker's lips, *Proc. of the 4rd EUROSPEECH Conference*, Madrid, Spain, 1995, pp. 1559-1562.

[2] ROBERT-RIBES, J., *Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique des voyelles*, Doctoral Thesis, INPG, Grenoble, France, 1995.

[3] SUMBY, W.H., & POLLACK, I., Visual contribution to speech intelligibility in noise, *J. Acoust. Soc. Am.*, 26, 1954, pp. 212-215.

[4] STORK, D., & HENNECKE, M., (Eds.), *Speechreading by humans and machines*, Springer-Verlag, Berlin, 1996.

[5] MARKEL, J.D., & GRAY, A.H.Jr., *Linear Prediction of Speech*, Springer-Verlag, New-York, 1976.

[6] LALLOUACHE, M.T., Un poste "visage-parole", *18th JEPs*, Montréal (Québec), Canada, 1990, pp. 282-286.