A SIGNAL PROCESSING SYSTEM FOR HAVING THE SOUND "POP-OUT" IN NOISE THANKS TO THE IMAGE OF THE SPEAKER'S LIPS: NEW ADVANCES USING MULTI-LAYER PERCEPTRONS

Girin L., Varin L., Feng G., Schwartz J.L.

Institut de la Communication Parlée, UPRESA 5009
INPG/ENSERG/Université Stendhal
B.P. 25, 38040 GRENOBLE CEDEX 09, FRANCE
E-mail: girin@icp.inpg.fr

ABSTRACT

This paper deals with the improvement of a noisy speech enhancement system based on the fusion of auditory and visual information. The system was presented in previous papers and implemented in the context of vowel to vowel and vowel to consonant transitions corrupted with white noise. Its principle consists in an analysis-enhancement-synthesis process based on a linear prediction (LP) model of the signal: the LP filter is enhanced thanks to associative tools that estimate LP cleaned parameters from both noisy audio and visual information.

The detailed structure of the system is reminded and we focus on the improvement that concerns precisely the associators: basic neural networks (multi-layers perceptrons) are used instead of linear regression. It is shown that in the context of VCV transitions corrupted with white noise, neural networks can improve the performances of the system in terms of intelligibility gain, distance measures and classification tests.

1. INTRODUCTION

It has been shown that there exists a complementarity between audition and vision for speech perception [9]. Thus, visual cues can compensate to a certain extent the deficiency of the auditory ones [12]. This explains that the fusion of auditory and visual information has met a great success in several speech applications, principally in speech recognition in noisy environments [1, 11].

We test here a slightly different idea, which is that the visual input could allow to *enhance* the audio input corrupted in acoustic noise. This idea has a theoretical basis. Recent work by Driver and colleagues suggest that the sensorial input in one modality can focus the attention of another modality on a specific part of its input. This has been demonstrated in [4] for visual-proprioceptive-tactile interactions, and in [3] for audio-visual interactions. In this last case, Driver presents various experiments in which the coordination of spatial attention across audition and vision enables the subjects to select sights and sounds from a common source in a selective listening task. These data suggest that audio-visual interactions could comprise a module enabling to perform what we propose to call "audio-visual scene analysis", in reference to the domain of "auditory scene analysis" [2] which focuses a great deal of interest in the field of audition.

In two previous papers [5, 6], we presented a new system dedicated to telecommunications or man-machine communication, in which we attempt to realize a first technical implementation of the idea that the acoustic signal of a given speaker could *pop out* in noise thanks to the visual input. In these first works, the audiovisual fusion/enhancement process was realized by a simple linear associator obtained by linear regression between noisy audiovisual data and clean audio data tuned from a learning corpus. The system was tested on vowel to vowel [5] and vowel to consonant (VCV) [6] single-speaker sequences and a quite good intelligibility gain was obtained on the vocalic parts of the signals, while the results on consonants were mitigated. This lead us to suspect the simplicity of the linear associator. In this paper, we present an improvement of the system with the use of non-linear associators for the fusion/estimation process. Thus, basic neural networks (multi-layers perceptrons) are used instead of the linear regression. New results obtained with both informal listening tests and objective measures (spectral distances, spectra classification) are presented.

2. STRUCTURE DESIGN

The system is essentially based on the linear prediction model (LP) [8] (fig. 1). First, an LP analysis is performed on the noisy signal. We obtain spectral parameters and the noisy speech residual signal is extracted by filtering through the inverse LP filter $A_n(z)$. Then, the noisy spectral parameters are combined with the video ones into an audiovisual vector so as to obtain estimated "cleaned" spectral parameters (see section 3). Finally, enhanced speech is synthesized by filtering the residual signal through the LP filter $1/A_e(z)$ derived from the "cleaned" parameters. The whole processing is performed frame-by-frame in the perspective of a continuous speech application.

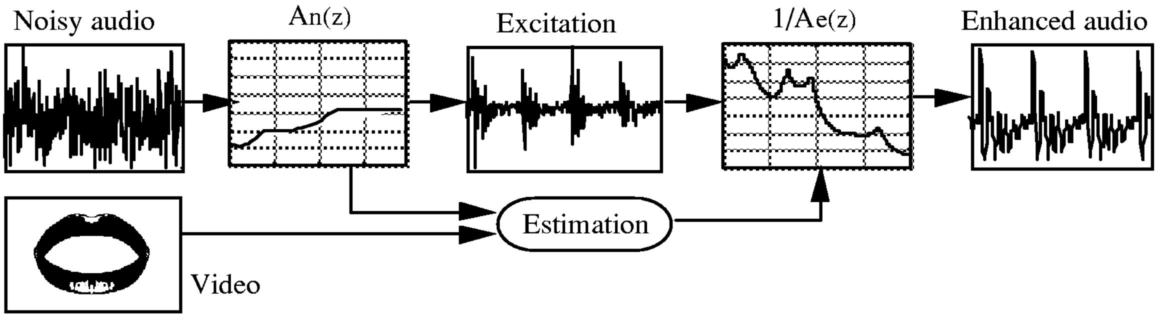

Figure 1: Structure of the noisy speech enhancement system

3. ASSOCIATORS

The estimation of the cleaned parameters is a classical "trained association" problem. In our previous works, linear regression was chosen for its simplicity and its efficiency. Its principle is simply to estimate each audio output parameter as a linear combination of the audiovisual input parameters. The matrix M of the coefficients of the linear combination are calculated by minimizing the mean square error $e=M_I.M-M_O$, where M_I and M_O are two matrix concatenating a large number of sets of respectively input and corresponding output parameters extracted from a training speech corpus.

Besides, neural networks have been widely used for classification tasks, like in the field of speech recognition, including recently audiovisual recognition [11]. They are theoretically able to approximate any non-linear function. That is why we choose to use them in order to improve the audio-visual association accuracy by enabling non-linearities between the input-output spaces. The networks used in this work are classical multi-layers perceptrons (MLP) based on error gradient backpropagation [10]. They have one hidden layer and the neuronal threshold functions are classical sigmoids. The backpropagation algorithm applied on the learning data is a very basic one, the only special tuning being the "momentum" option: the adjustment of each weight takes in account its previous modification during the training iterations thanks to a forgetting coefficient.

4. EXPERIMENTATION

4.1. Video and audio inputs

Video stimuli consist in three basic geometric parameters describing the speaker's lip shape, namely internal width (A), height (B) and area (S) of the labial contour. These parameters are automatically extracted every 20 ms thanks to the ICP face processing system [7].

Concerning the audio LP parameters, 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 synchronize with the 20 ms video period). To obtain the filter $1/A_e(z)$ from the "cleaned" spectral parameters, we process an inverse FFT on the squared linearly-scaled coefficients, and apply a 20-order Levinson procedure on the resulting estimated autocorrelation coefficients [8]

4.2. The corpus

The corpus, already used in [6], consists in $V_1CV_2CV_1$ sequences uttered by one speaker. V_1 and V_2 are within [a, i, y, u]. C is within the plosives set [p, t, k, b, d, g]. One item of each of the 96 possible stimuli (4x4x6) is used during the training of the associators 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 2500 audiovisual vectors for a series of 96 stimuli (about 25 frames by stimuli).

4.3. Experimental protocol

We consider only the case of an additive white noise. In order to sufficiently generalize the process with respect to the noise level, the training of the associators is done with input audio corrupted at different levels. The results discussed here are obtained with the use of two different learning/processing conditions for both linear and neuronal associators. In the first one, the stimuli frames are presented at signal to noise ratios (SNRs) of ∞, 18, 12, 6 and 0 dB. In the other one, the stimuli frames are 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 categorization in the "small" or "strong" noise 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. Between 0 and 6 dB, the two "small/strong noise" associators provide quite similar outputs.

5. EXPERIMENTAL RESULTS

5.1. Informal listening tests

In [5], it was observed that the linear associator allowed a significant enhancement of vocalic transitions on a large SNRs scale (from 18 to –18 dB). In [6], where the VCV corpus was used, this efficiency was somehow reported on the vocalic parts of the stimuli while all consonants stayed poorly intelligible. New informal listening tests using the MLP reveal that this associator seems to allow a better global enhancement than the linear associator at any level of noise: when listening successively two occurences of a stimuli provided by both associators, the one enhanced by the MLP is systematically preferred. This seems to be due first to an improvement of the enhancement of vocalic sections. Besides, a significant improvement on consonants is obtained for the [p, b] sections, showing that the non-linear relation between closed lip shapes and acoustic features can be exploited by the neural network. Unfortunately, other consonants are not significantly enhanced (even sometimes degraded, which was already observed with the linear associator), showing anew the difficulty to exploit poorly visible articulatory gestures.

5.2. Distance measures

The Itakura distance [8] has already been used in [5, 6] to measure the difference between the enhanced and clean spectra. Rather small distances were obtained compared to those between noisy and clean spectra. Hence, the procedure does produce a significant enhancement and the new plot in figure 2 allows to compare the results obtained with the linear associator and the MLP for two values of the hidden layer neurons number (40 and 200). The distances are calculated and averaged on the complete test corpus (96 stimuli) for 8 different SNRs (∞, 18, 12, 0, -6, -12, -18 dB).

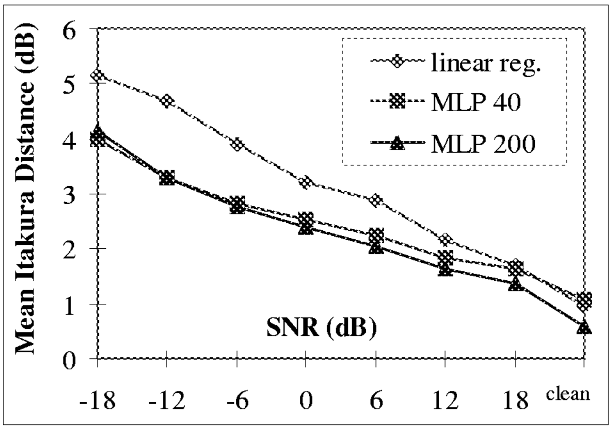

Figure 2: Mean Itakura distance between enhanced and clean spectra of the test corpus for the linear regression and MLPs with 40 and 200 neurons.

In each case, that is 40 or 200 neurons, the performances of the MLP significantly overcome the linear associator ones for strong levels of noise (until SNR = 6 dB). Below 0 dB, the results seem independent of the number of neurons. This is confirmed by figure 3, which displays the same distances as a function of the number N of neurons of the hidden layer. At smaller levels of noise, the gain gets smaller if N is small (it is even negative for 20 neurons at ∞ or 18 dB). A quite larger number of neurons seems necessary to obtain better results than the linear regression at small levels of noise.

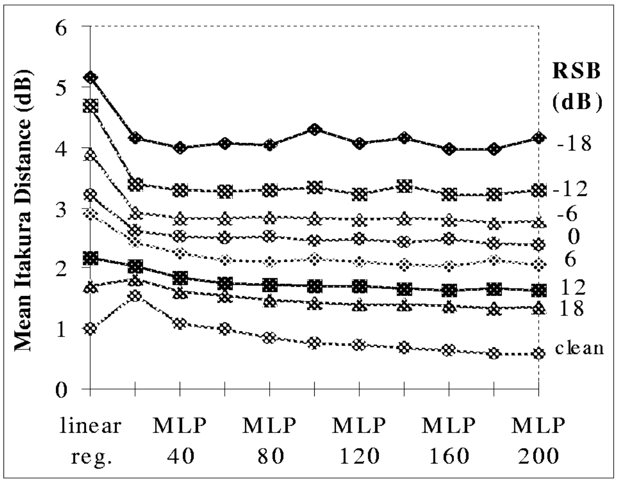

Figure 3: Mean Itakura distance between enhanced and clean spectra of the test corpus for the linear regression and MLPs as a function of the number of neurons of the hidden layer.

As a summary, a relatively small number N of neurons with respect to data size allows a quite more efficient output estimation than the linear regression in terms of spectral distances. This is obtained on a large noise level scale (below 6/12 dB). As N increases, this gain of performances is spread to small noise levels.

5.3. Gaussian classification test

To evaluate the system in a recognition task, gaussian classification tests have been realized separately on the 4 vowels and the 6 consonants of our stimuli. The items used in this test are two selected frames near each vocalic nucleus of each stimulus for the vowels, and near each burst for the consonants. For each level of noise, we obtain 576 vowel and 384 consonant items (2 selected frames, 3 vowels and 2 consonants occurrences per stimulus, 96 stimuli), that is to say 144 per vowel, and 64 per consonant. Since the number of data is small compared to the number of audio parameters, we reduce it from 50 to 10 by means of a principal components analysis (PCA). Both the PCA and the gaussian audio classifier parameters are determined with learning data presented at 3 levels of noise (∞, 18, 12 dB). Figure 4 displays the classification scores obtained separately on vowels and consonants for three conditions: noisy audio, audio enhanced with the linear regression, and audio enhanced with the MLP. In this last case, the number of neurons is chosen not too great (120) while allowing the best global scores. All scores are normalized between 0 and 100%, with 0% corresponding to a random choice and 100% to perfect recognition.

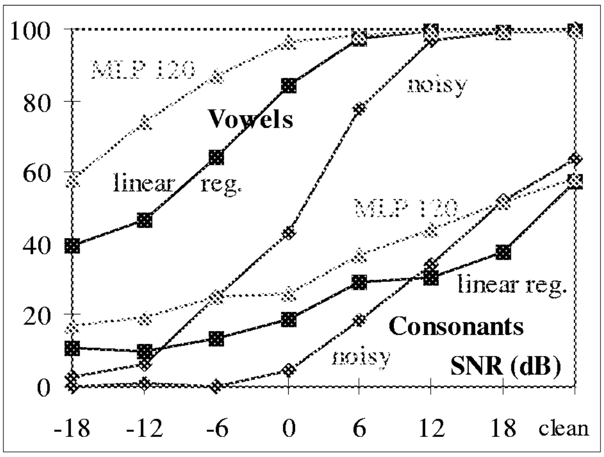

Figure 4: Gaussian classification test scores

The difference of the scores in the noisy and enhanced conditions confirms the efficiency of the system. The gain of the MLP with respect to the linear associator is anew notable. In particular, the network is able to clearly improve the "reshaping" of the vowel spectra, while the regression was shown to be already efficient in that case [5, 6]. The scores obtained on consonants are less impressive, even if a gain is also observed from linear regression to the MLP. All these results confirm the observations of section 5.1.

6. CONCLUSION

We have presented in this paper the improvement of an original noisy speech enhancement system introduced in [5]. This system is based on a filtering process exploiting some fusion/estimation process from both auditory information and lip contour parameters. The improvement concerns the use of neural networks (multi-layer perceptrons) for the fusion/estimation process, instead of linear regression which was used in [5] to show the feasability of the study.

It has been shown that in the context of VCV transitions corrupted with white noise, neural networks can improve the performances of the system in terms of intelligibility gain, distances measures and classification tests. The gains observed on consonants are mitigated: only the [p b] lip closures seem to be exploited correctly, which is a new (even not to much surprising) result compared to previous one obtained with the linear associator. Hence, the improvement of consonants stays the key-point of this work. The good surprise comes from the vowels: their enhancement is significantly improved by the MLP although the performances of the linear regression for the vowels were already quite good.

To compare with [5], formal perceptive tests remains to be done. This is part of our future works, as much as the association of such audiovisual methods with classical acoustic ones (e.g. multi-microphones systems).

7. REFERENCES

1. Adjoudani, A., and Benoît, C. "On the integration of auditory and visual parameters in an HMM-based ASR", in [11], pp. 461-472.
2. Bregman, A.S., *Auditory Scene Analysis: the perceptual organization of sound*, Bradford Books, MIT Press, Cambridge, Mass, 1990.
3. Driver, J. (1996), "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading", Nature, 381, 66-68.
4. Driver, J., and Grossenbacher, P.G., "Multimodal spatial constraints on tactile selective attention". In Inui, T., and McClelland, J.L. (Eds.), *Attention and Performance XVI: Information Integration in Perception and Communication* (pp. 209-236), Bradford Books, MIT Press, Cambridge, Mass, 1996.
5. Girin L., Feng G., and Schwartz J.-L., "Noisy speech enhancement by fusion of auditory and visual information : a study of vowel transitions", *Proc. of the 5rd Euro. Conf. on Speech Communication and Technology*, Rhodes, Greece, pp. 2555-2558, 1997.
6. Girin L., Feng G., and Schwartz J.-L., "Can the visual input make the audio signal "pop out" in noise? A first study of the enhancement of noisy VCV acoustic sequences by audiovisual fusion", *Proc. of the 1rst Euro. Tutorial & Research Workshop on Audio-Visual Speech Processing*, Rhodes, Greece, pp.37-40, 1997.
7. Lallouache M.T., "Un poste visage-parole", *18èmes Journées d'Etudes sur la Parole*, Montréal, Québec, p. 282-286, 1990.
8. Markel, J.D., and Gray, A.H.Jr., *Linear Prediction of Speech*, Springer-Verlag, New-York, 1976.
9. Robert-Ribes, J., Piquemal, M., Schwartz, J.L., and Escudier, P., "Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition", in [11], pp. 193-210.
10. Rumelhart D.E., Hinton G.E. and Williams, R.J., "Learning representations by back-propagating errors", *Nature*, 323, pp. 533-536, 1986.
11. Stork, D.G., and Hennecke, M.E. (Eds.), *Speechreading by humans and machines: theories, models and applications*, Springer-Verlag, Berlin, 1996.
12. Sumby W.H., and Pollack I., "Visual contribution to speech intelligibility in noise", J. Acoust. Soc. Am., 26, pp. 212-215, 1954.