LINEAR EQUALIZATION OF THE MODULATION SPECTRA: A NOVEL APPROACH FOR NOISY SPEECH RECOGNITION

F. Díaz-de-María, J. Vicente-Peña, A. Gallardo-Antolín and C. Peláez-Moreno

Signal Theory and Communications Department, EPS-Universidad Carlos III de Madrid, Spain

fdiaz, jvicente, gallardo and carmen@tsc.uc3m.es

ABSTRACT

In this paper, we have tackled the problem of noisy speech recognition. In particular, we have presented a novel approach to the design of filters for processing the modulation spectrum, that we have called linear equalization. We postulate that, as long as the distortion of the spectral parameters due to noise can be modeled as linear, an advantageous solution consists on estimating this linear perturbation system and designing its inverse system (the equalizer). Our experimental results show that the proposed method is very effective for three of the five considered noises.

1. INTRODUCTION

In real world applications, Automatic Speech Recognition (ASR) systems often encounter situations in which there is a mismatch between training and testing conditions (e.g. noise, transmission channel). In such scenarios, there is a dramatic degradation of the recognizer accuracy.

During the past years, a family of techniques has been proposed for dealing with this type of problems, such as robust parameterizations, feature vector adaptation and model parameter compensation. In this paper, we have focused on the first approach, i.e., extracting robust speech features that are relatively insensitive to different sources of noise.

For that purpose, it would be interesting that the front-end were able to keep the linguistic information contained in the speech signal and reject the information related to noise. Perceptual experiments show that the intelligibility of speech is mostly concentrated in some bands of the modulation spectrum, while the rest do not seem to contribute importantly. According to Kanedera et al. [1]:

- In clean environments, most of the useful information is contained in the frequency range between 1 and 16 Hz of the modulation spectrum.
- The band around 4 Hz (it would roughly correspond to syllabic rates) is the most useful component in both, clean and noisy conditions.
- In noisy environments, the components of the modulation spectrum below 2 Hz and above 10 Hz are less important for speech intelligibility. In particular, the band below 1 Hz contains mostly information about the environment (e.g. the effects introduced by the frequency characteristic of the transmission channel). Therefore, the recognition performance can be improved by suppressing this band in the parameterization process.

Typically, that suppression can be performed by temporal filtering of time trajectories in the logarithmic spectrum or cepstral domain. For example, CMN ("Cepstral Mean Normalization") [4] is a high-pass filter, which eliminates the DC component of the cepstrum parameters. The classical derivative features (delta) [5] can be seen as a filtering of the static parameters, in which the components around 10 Hz are enhanced. RelAtive SpecTral technique (RASTA) [2] is a band-pass filter, which keeps the frequencies belonging to the frequency range between 1 and 12 Hz. In addition, other more complex filters (Slepian filters) [3] have been proposed for convolutional noise conditions (specifically, telephone environment).

In this paper we propose a novel method for designing the modulation spectrum filters that we call *linear equalization* and we asses our method in scenarios with several additive noises.

The paper is organized as follows. The linear equalization approach motivation is described in section 2. Section 3 is devoted to the experimental assessment of the proposed method and finally, we draw some conclusions and outline future work in Section 4.

2.LINEAR EQUALIZERS FOR MODULATION SPECTRA FILTERING: CONCEPT AND DESIGN

2.1. The Linear Equalization Approach

Most of the previous filtering approaches aiming at selecting or enhancing the frequency band of the modulation spectrum responsible for the intelligibility use linear filtering. In this context, our statement of the problem is as follows. We assume, as implicitly others do when using linear filters, that the distortion of the spectral parameters due to noise can be modelled as linear. Thus, we go one step further and try to design the inverse linear filter we have called a *modulation spectrum equalizer*. A more detailed discussion of our proposal follows.

Figure 1 schematically shows the proposed method for computing estimates of the modulation spectra. Firstly, P

sequences of the spectral parameters are obtained from clean speech, being P the number of parameters conforming the speech recognition feature vector. In particular, we use 12 Mel-Frequency Cepstral Coefficients (MFCC), $\left\{c_s^i[n],\ i=1,\ldots,P\right\}$ –where the sub-index s stands for (clean) *speech* as opposed to *ns* used for *noisy speech*–. Finally, we calculate estimates of the spectrum of each spectral parameter sequence, thus obtaining the so called estimated modulation spectra, $\left\{\hat{H}_s^i(\Omega),\ i=1,\ldots,P\right\}$.

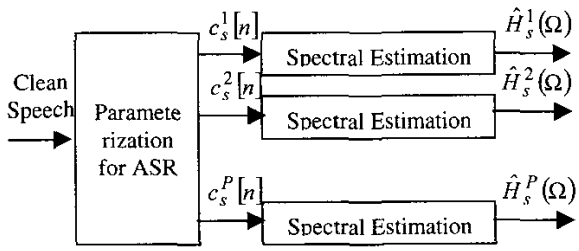

Figure 1.- Computattion of the modulation spectra

When dealing with noisy speech, noisy MFCCs, $\left\{c_{ns}^i[n],\ i=1,\ldots,P\right\}$, and their corresponding modulation spectra $\left\{\hat{H}_{ns}^i(\Omega),\ i=1,\ldots,P\right\}$ are also obtained in the same manner.

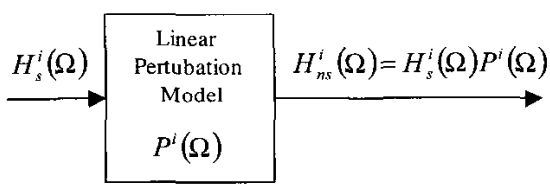

Figure 2.- Linear Perturbation Model

If, as illustrated in Figure 2, it is assumed that noise causes a linear distortion, $\left\{P^i(\Omega),\ i=1,\ldots,P\right\}$ then, by designing the inverse filter,

$$\left\{P_{inv}^i(\Omega)=\frac{1}{P^i(\Omega)}=\frac{H_s^i(\Omega)}{H_{ns}^i(\Omega)},\ i=1,\ldots,P\right\} \qquad (1)$$

we would be able to compensate it.

From our point of view, this would be the best solution as long as the distortion is linear. Though, obviously, most of the times, the distortion is not linear and thus only some part of it could be linearly modelled, this approach is the best thing attainable by linear filtering.

2.2. Equalization Filters Design

For our experiments, we have used two parameterization procedures. 1) MFCC directly computed from the speech signal; and 2) MFCC derived from a LP spectrum (henceforth, LP-MFCC). It is well known that some type of spectral smoothing (in our case a LP spectrum) turns out to be beneficial in noisy environments [8]. More details about both parameterization methods will be supplied in Section 3.

Therefore, one filter per spectral parameter, type of noise and parameterization method has been designed. These filters should remove the contribution to modulation spectra due to noise. For their design, we have used the estimated the mean spectrum of each individual spectral parameter time series corresponding in two situations: 1) clean speech, $\left\{\hat{H}_s^i(\Omega),\ i=1,\ldots,P\right\}$, and 2) noisy speech, $\left\{\hat{H}_{ns}^i(\Omega),\ i=1,\ldots,P\right\}$.

Finally, the frequency response of ideal inverse filters, $\left\{P_{inv}^i(\Omega),\ i=1,\ldots,P\right\}$ could be calculated using (**1**).

To be precise, the proposed method can be implemented following two main steps:

1) *Computation of estimations of the modulation spectra* for both, clean and noisy speech, i.e.,

$$\left\{\hat{H}_s^i(\Omega)\ and\ \hat{H}_{ns}^i(\Omega),\ i=1,\ldots,P\right\}.$$

For that purpose, either P MFCCs or P LP-MFCCs must be extracted ($P=12$) where the sampling frequency of the cepstral time series we have employed is 10 ms. Therefore, every 10 ms. we obtain P energy-density spectrum estimates,

$$\left\{\left|\hat{H}_s^i(\Omega)\right|^2\ and\ \left|\hat{H}_{ns}^i(\Omega)\right|^2,\ i=1,\ldots,P\right\}.$$

Finally, we average these energy-density spectra obtaining:

$$\left\{\left|\hat{H}_s^i(\Omega)\right|_{av}^2\ and\ \left|\hat{H}_{ns}^i(\Omega)\right|_{av}^2,\ i=1,\ldots,P\right\}$$

2) *Filter design*: As already mentioned, we have designed one filter per parameter. Each filter has been designed to approximate $\left\{P_{inv}^i(\Omega),\ i=1,\ldots,P\right\}$ as follows:

$$\left\{\hat{P}_{inv}^i(\Omega)=\sqrt{\frac{\left|\hat{H}_s^i(\Omega)\right|_{av}^2}{\left|\hat{H}_{ns}^i(\Omega)\right|_{av}^2}},\ i=1,\ldots,P\right\}$$

We have used IIR filters with 4 poles and 1 zero. Figure 3 illustrates one example of such filter approximation.

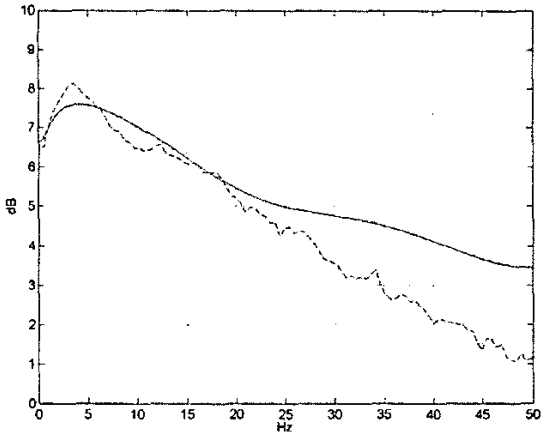

Figure 3 An example of filter approximation. The dashed line represents the actual $\left\{\hat{P}^i_{inv}(\Omega),\ i=1,...,P\right\}$ and the solid line the corresponding IIR approximation.

3. EXPERIMENTS AND RESULTS

3.1. Databases and Baseline Systems

The database employed in our experiments is the well-known Resource Management RM1 Database [6], which has a vocabulary of 991 words. The training corpus consists of 3990 sentences and the test set contains 1200 sentences, which corresponds to a compilation of the first four official test sets. We have used a downsampled version (at 8 KHz) of the database (originally recorded at 16 kHz in clean conditions), context-dependent acoustic models (three-mixture cross-word triphones) and a simple language model (a word-pair grammar).

3.2. Adding Noises

Five different types of noises (pink, white, babble, factory and Volvo) from the NOISEX database [9] are added to the speech signal to achieve a signal to noise ratio of 12 dB. As we have used clean speech for estimating the acoustic models, the noises are only added for testing the recognition performance.

3.3. Parameterization

Here, we summarize the two alternative ways to obtain the parametric representation of the speech signal from which the recognition is performed:

MFCC: 12 mel-cepstral and a log-energy coefficients are extracted every 10 ms using a hamming window of 25 ms from the speech. Each individual MFCC coefficient is filtered using the IIR designed filters. Finally, 12 delta-cepstra and a delta log-energy coefficients are appended.

LP-MFCC: 10 LPC (*–Linear Prediction Coefficients–*) and an energy coefficient are firstly computed from speech at the same rate than the previous parameterization. These parameters are transformed into 12 MFCC plus energy being the rest of the procedure the same that for the previously described parameterization.

3.4. Confidence Measures

In order to state the statistical significance of the experimental results we have calculated the confidence intervals (for a confidence of 95%) using the following formula [7], (pp. 407-408):

$$\frac{band}{2} = 1.96\sqrt{\frac{p(100-p)}{n}}$$

where p is the word accuracy and n is the number of examples to be recognized (10,288 words). Thus, any recognition rate will be presented as belonging to the band $\left[p-\frac{band}{2}, p+\frac{band}{2}\right]$ with a confidence of 95%.

3.5. Results

3.5.1 Filters Design

As previously mentioned, we have designed one filter per coefficient and per noise. Nevertheless, observing the frequency response of the equalizers, it can be concluded that the designed filters are quite independent of the considered type of noise.

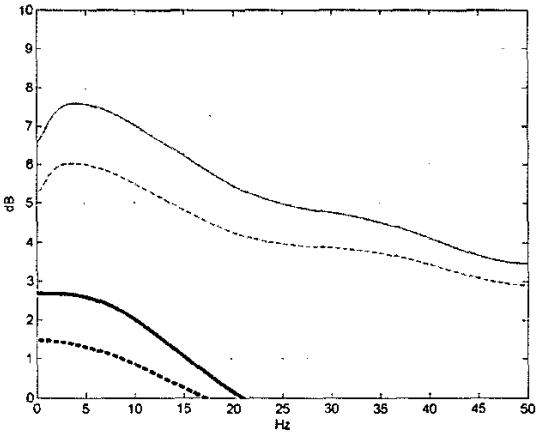

Figure 4.- Equalizers obtained for MFCCs #1 (thin lines), and #12 (thick) and two types of noises, factory (solid line) and babble (dashed line)

Figure 4 illustrates this observation for four different MFCCs (the first and the last ones) and two types of noises (factory and bable).

3.5.2 Recognition Scores

Table 1 summarizes all the recognition results for the five considered type of noises. Confidence intervals appear in brackets below each recognition score. The first and second row show the baseline experiments with the

considered parameterizations, MFCC and LP-MFCC, respectively. Our first conclusion is clear: except for the white noise, computing the MFCC from the LP spectrum provides very significant improvements. As previously indicated, these are well-known results.

Noise	Pink	Volvo	Factory	White	Babble
MFCC	25.39 (24.5, 26.3)	70.39 (69.5, 71.3)	29.24 (28.3, 30.2)	12.74 (12.0, 13.4)	22.01 (21.2, 22.9)
LP-MFCC	38.04 (37.1, 39.0)	76.65 (75.8, 77.5)	36.72 (35.8, 37.7)	8.57 (8.0, 9.2)	29.56 (28.6, 30.5)
Equalized MFCC	30.63 (29.7, 31.6)	67.02 (66.1, 68.0)	33.94 (33.0, 34.9)	16.82 (16.0, 17.6)	25.29 (24.4, 26.2)
Equalized LP-MFCC	42.33 (41.3, 43.3)	75.09 (74.2, 76.0)	41.22 (40.2, 42.2)	13.43 (12.7, 14.1)	29.51 (28.6, 30.4)

Table 1.- ASR results for several types of noises

The third and four rows show, for MFCC and LP-MFCC, respectively, the results achieved by our proposal. As it can be seen, we obtained significant improvements for three of the five considered noises, namely, *pink*, *factory* and *white*. Focusing on the LP-MFCC parameterization, the proposed technique achieves improvements of 11.3 %, 12.2 % and 56.7 % for *pink*, *factory* and *white* noises, respectively.

Considering the achieved results and provided that our method is designed to deal with linear perturbations, it can be concluded that the degree of linearity of these perturbations notably depend on the type of noise being considered. On the one hand, *pink*, *factory* and *white* perturbations exhibit some relevant linear component (since our method turns out to be effective). On the other, perturbations due to *Volvo* or *babble* noises should be mainly non-linear.

4. CONCLUSIONS AND FURTHER WORK

In this paper, we have tackled the problem of noisy speech recognition (additive noise). In particular, we have focused on those robust techniques based on filtering of temporal trajectory of the spectral parameters. In this context, we have introduced a novel approach for the design of filters for processing the modulation spectrum, that we have called linear equalization.

We postulate that, as long as the distortion of the spectral parameters due to noise can be modelled as linear, the best solution consists on estimating this linear perturbation system and designing its inverse system (the equalizer).

Our experimental results show how the proposed method is very effective for three of the five evaluated noises. Therefore, it can be concluded that the degree of linearity of the perturbation (and consequently the effectiveness of our method) notably depends on the type of noise considered.

Currently, we are working on the design of a filter bank useful for all of the noises. In parallel, we are computing new reference results using Lin-log RASTA (one of the most well-known filtering-based techniques). Preliminary results indicate that RASTA performance is quite below those achieved by equalization.

5. ACKNOWLEGMENTS

This work has been partially supported by Spanish Regional grant CAM-07T-0018-2000.

6. REFERENCES

[1] N. Kanedera, H, Hermansky y T. Arai, *"On Properties of Modulation Spectrum for Robust Automatic Speech Recognition"*, Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing: ICASSP 1998, vol. 2, pp. 613-616, Seattle (USA), 1998.

[2] H. Hermansky y N. Morgan, *"RASTA Processing of Speech"*, IEEE Transactions on Speech and Audio Processing, vol. 2, nº 4, pp. 578-589, 1994.

[3] C. Nadeu, P. Pachès-Leal y B. H. Juang, *"Filtering the Time Sequences of Spectral Parameters for Speech Recognition"*, Speech Communication, 22, pp. 315-332, 1997.

[4] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", IEEE Transactions Acoustics, Speech and Signal Processing, vol. 29, pp. 254-272, 1981.

[5] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", IEEE Transactions Acoustics, Speech and Signal Processing., vol. 34 (1), pp. 52-59, 1986.

[6] National Institute of Standards and Technology (NIST) (distributor): The Resource Management corpus part 1 (RM1) (1992)

[7] Weiss, N.A., Hasset, M.J., "Introductory statistics", Third Edition. Reading, MA: Addison-Wesley, 1993.

[8] Gold, B., Morgan, N., "Speech and Audio Signal Processing", New York, NY: John Wiley & Sons, 2000.

[9] A. P. Varga, J. M. Steenneken, M. Tomlinson y D. Jones. "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition."Tech. Rep. DRA Speech Res. Unit. Malvern, Worcestershire, U. K. 1992.