EQUIVALENCES AND LIMITS OF PITCH-BASED TECHNIQUES FOR ROBUST SPEECH RECOGNITION

Juan A. Morales-Cordovilla, Antonio M. Peinado and Victoria Sánchez

Dept. of Teoría de la Señal Telemática y Comunicaciones, Universidad de Granada, Spain, {jamc,amp,victoria}@ugr.es.

ABSTRACT

This paper discusses the performance limits of pitch-based techniques which, in some way, use the pitch in order to carry out robust ASR (Automatic Speech Recognition) under noise conditions and which employ minimal assumptions about the noise. In order to do so, we will identify the basic robust mechanisms employed by these techniques for recognizing voiced frames, the optimum mechanisms will be identified (by means of some equivalences), and the corresponding limit results will be experimentally obtained by applying MD oracle masks and ideal pitch. Experimental results with Aurora-2 database will show that *Pitch-based Noise Estimation* [10] for MD recognition is close to the limits of the pitch-based robust ASR techniques, although it would require additional information in order to achieve the performance with MD oracle masks.

Index Terms— Robust speech recognition, pitch, noise estimation, limits, missing data, harmonic tunnelling.

1. INTRODUCTION

Acoustic noise represents one of the major challenges for ASR (Automatic Speech Recognition) systems. Many different approaches have been proposed to deal with this problem in monaural signal [13, 7, 17] and many of them try to employ some kind of noise information to do robust ASR. However, when one wants to deal with all kind of noises it is clear that the most important information to separate noise from speech is just speech information. There exits many cues and informations which help to distinguish speech from noise but at the end the correct choice will depend on what is defined as speech. Speech can be emitted in many different ways which mainly depend on the considered type of the "‘main source"’. These ways can be whispering, vocal harmony speech (in music), etc.. In this paper it will be considered that speech is emitted in its normal way, with vibration of the vocal folds and with only one pitch at each time instant.

Continuing with the search for the most important cues, this paper will particularly consider the signal pitch due to the three following reasons:

1. Many psychoacoustics experiments, such as those shown in [4, 17], reach the conclusion that very often humans use pitch to separate speech from noise.

2. Pitch is a useful information to distinguish different types of speech segments (voiced, unvoiced and silence) and to separate speech and noise signals.

3. Many robust ASR techniques inspired in human recognition, as shown in [17], use pitch.

The comparison of the different ASR techniques based on pitch (found in the bibliography) is not an easy matter because of several reasons:

1. Each author uses a different pitch extractor to evaluate his technique.

2. It is not clear which is the real cause for obtaining different results: different mechanisms applied to voiced and unvoiced sounds, application of additional techniques (such as cepstral normalization, missing data approaches,...), etc.

3. Sometimes it is not clear whether an author is proposing either a new technique for robust ASR based on pitch or a new robust pitch extractor (or both at the same time).

RTH 2011 v.. This work has been supported by the Spanish MEC/FEDER project TEC2010-18009

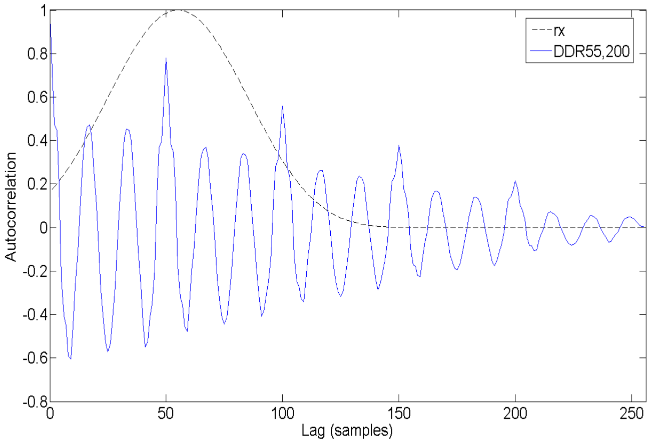

Fig. 1. Example of $DDR_{55,200}$ window applied to the OSA of a voiced frame with a pitch value of 50 samples.

Because of these reasons, we consider it necessary to do a fair comparison of these pitch-based techniques, trying to show the equivalences between some of them and trying to see the limits of pitch-based recognition but without considering pitch extraction.

The structure of the paper is as follows. First, a classification of the pitch-based techniques based on the robust mechanism applied to voiced frames is presented. Section 3 explains the different mechanisms applied to the other frames (silence and unvoiced frames). Section 4 investigates about which is the best mechanism for recognizing voiced frames. In section 5, the performance limit of pitch-based techniques will be experimentally shown on the Aurora-2 database. The paper concludes with a summary and a discussion of future works.

2. VOICED MECHANISMS

In principle, pitch-based techniques can be supposed as different if we only pay attention to some specific details (pitch extractor, processing of unvoiced and silence frames, etc.). However, they can be reduced to one of following four basic mechanisms which depend on the robust method applied to voiced frames: exploitation of the harmonic structure, comb estimation of clean signal, tunnelling estimation of noise and harmonicity mask estimation.

2.1. Exploitation of harmonic structure

These mechanisms do not require a pitch extraction but only some properties which can be derived from periodicity. For example, *Asymmetric Window* [12] proposes to employ a so-called $DDR_{55,200}$ window on the OSA (One Side Autocorrelation) of voiced frames in order to extract a more robust spectrum and then AMFCC (Autocorrelation Mel Frequency Cepstral Coefficients) features. Fig. 1 shows an example of the $DDR_{55,200}$ applied to the OSA of a voiced frame with pitch 50 samples. 55 is the center and 200 is the width of the window. The election of this proposed window is based on two considerations: 1) The autocorrelation coefficients which are less affected by the noise are those placed at pitch (or its multiples) lags, this the reason of centering the window on 55 samples (145 Hz at a 8000 Hz sample frequency) which is the mean human speech pitch. 2) A wide of 200 samples provides few weight to the first autocorrelation coefficients which usually are more affected by noise (mainly if the noise is white-like or contained in the first autocorrelation lags).

Other related techniques are SWP [9] and HASE [16] which try to ”‘clean”’ the signal using these properties, and HF [14] which estimates the noise by exploiting the spectral harmonic shape.

2.2. Comb estimation of clean signal

These mechanisms use the pitch frame to apply some kind of comb filtering, i. e. some kind of algorithm which can be reduced to a sort of removing noise between the gaps (or tunnels) which are in the middle between the pitch spectrum harmonics. The resulting clean signal can be recognized from its cepstral representation.

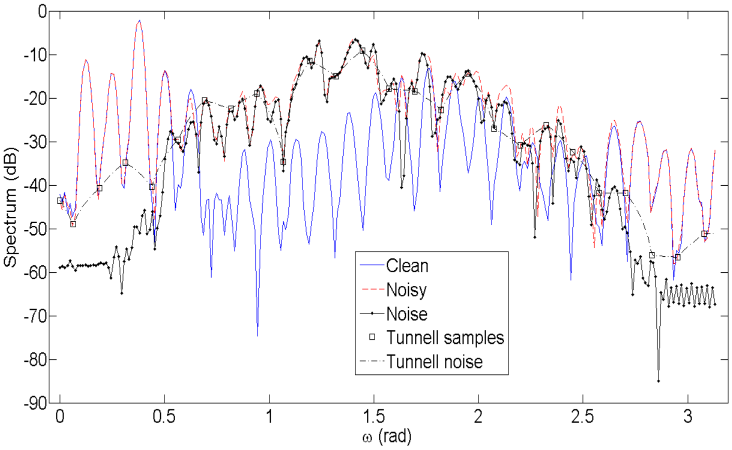

Fig. 2. Example of tunnelling noise estimation on a voiced noisy frame with pitch $\omega_0 = 0.126$ rad..

For example, *Sifting autocorrelation* [11] uses the pitch to obtain a clean autocorrelation estimate of voiced frames, the clean spectrum estimate and the AMFCC feature. This mechanism depends on a sifting parameter δ which informs about the amount of autocorrelation products which are rejected because they are supposed to be more contaminated by white-like noises. Sifting has the advantages of the comb techniques (eliminating the noise placed between pitch harmonics) and Asymmetric Windows (eliminating white-like noises). $delta = 8$ is taken as the best value for Aurora-2 database.

Other related techniques are WHNM [15] and PHCC [6].

2.3. Tunnelling estimation of noise

These mechanisms are the opposite of the preceding ones and estimate noise spectrum (tunnelling noise) employing tunnelling samples, that is, the spectral gaps between the harmonics. The resulting noise estimate can be employed in SS, MD, etc.. For example, in *Pitch-based Noise Estimation* [10] the discrete tunnelling noise is estimated by interpolating tunnelling samples which are obtained from continuous noisy spectrum by means of the pitch frequency. Fig. 2 shows an example of tunnelling noise estimation, it can be observed that only one tunnelling sample (between the harmonics) is taken.

This noise estimate could be employed in a SS (Spectral Subtraction) approach but the author proposes to use it in a MD (Missing Data) HMM-recognizer [3] because MD does not need to estimate perfectly the noise level but only where speech is dominated by noise (i. e. MD mask). Soft mask (between $[0, 1]$) of mel-noisy spectrum $M_y(f)$ is obtained by passing the local SNR estimate through a sigmoid function (which depends on a threshold and a slope). This local SNR can be obtained by means of the mel-tunnelling noise $M_{\hat{N}}(f)$ estimate as follows:

$$SNR(f) = 20log_{10}\frac{|M_Y(f) - M_{\hat{N}|}(f)}{M_{\hat{N}}(f)} \tag{1}$$

Other related techniques are HT [5] (which estimates noise from discrete spectrum) and FPM-NE [2] (which estimates noise by employing a temporal filter of the type of $h_T(t) = \delta(t) - \delta(t - T)$ where T is the pitch period in samples).

2.4. Harmonicity mask estimation

This mechanism estimates the mask of each frequency-temporal pixel by means of the correlogram and the pitch. Techniques which use cochleagram such as that of Barker's [1] employ this mechanism in which the soft-mask of noisy spectrum is estimated by passing harmonicity of each output cochleagram channel f through a sigmoid function. This harmonicity is obtained by means of the correlogram $A_y(f, k)$ as follows:

$$H(f) = A_y(f, T)/A_y(f, 0) \tag{2}$$

where T is the pitch of the frame and $A_y(f, 0)$ is the 0-lag autocorrelation coefficient of channel f. In general, ASA (Auditory Scene Analysis) techniques such as those presented in [17, 8] employ the harmonicity to estimate the mask.

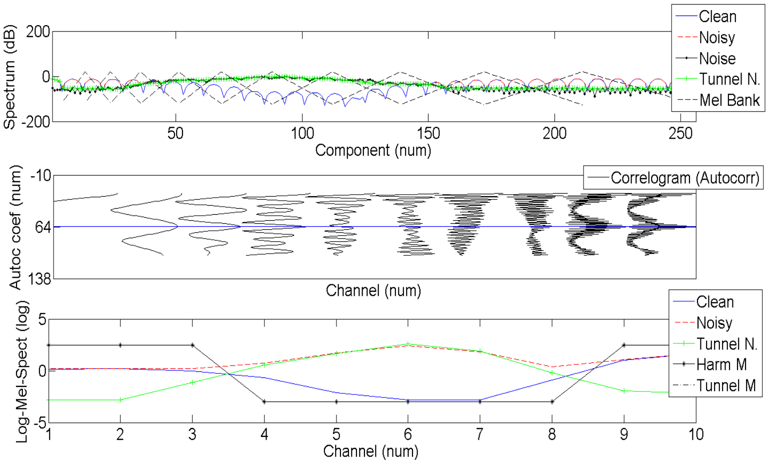

Fig. 3. Comparison of the mechanisms to estimate a tunnelling mask and a harmonicity mask. Both masks are shown in the Log-Mel Spectrum plot

Taking into account these four mechanisms we can investigate about which is the best one and whether they fully exploit the pitch information to improve the recognition in voiced frames. These questions will be answered in Sec. 4.

3. MECHANISMS FOR SILENCE AND UNVOICED FRAMES

Regarding mechanisms applied to the silence and unvoiced frames we can say that each technique applies its own mechanism. For example, in Asymmetric Window and Sifting, the same mechanism of voiced frames is also applied to silence and unvoiced frames (by means of a constant pitch fictitious in the case of Sifting). For silence frames, it is clear that the application of these mechanisms are always beneficial, but the application to unvoiced frames could degrade the performance (mainly at clean conditions) because the information of unvoiced sounds is mainly contained in the first autocorrelation coefficients, which tend to be removed. Nevertheless, this problem can be avoided by applying the same technique in both, training and test stages.

On the other hand, Pitch-based Noise Estimation employs a VAD (Voice Activity Detector) to estimate the mask of unvoiced frames by extrapolating the noise from silence regions. Original Barker's technique uses 10 first noisy frames to estimate the noise in silence and unvoiced frames and then obtain the mask (by mean of local SNR estimation) in these frames. In this paper an adaptation of Barker's technique, which use the same VAD and similar unvoiced mask estimation than Pitch-based Noise Estimation, will be developed in order to do a fair comparison of both techniques.

4. OPTIMUM VOICED MECHANISMS

4.1. Comparing tunnelling and harmonicity masks

It can be shown that the mask derived from tunnelling noise is similar to that derived from harmonicity measures if similar channel numbers and a suitable selection of thresholds are applied.

Fig. 3 can help to understand this similarity. The clean and tunnelling noise estimate, which indicates where the mask should be 1 or 0, are on top of the picture along with the 10 Mel filter bank, employed in tunnelling estimation. The outputs of the 10 gammatone channels of the correlogram employed to estimate harmonicity mask are in the middle plot. The two mask estimates (*Harmonicity and Tunnelling Mask*) are overlapped at the bottom of the picture along with the Log-Mel spectra employed to estimate the tunnelling mask, showing the strong similarity of both estimates. We can conjecture that both masks will yield similar recognition results (hypothesis H1).

4.2. Optimum pitch-based noise estimation

Let's suppose that we have a noisy signal $x(n)$ of length N which is the sum of a pure periodic clean signal $p(n)$ and a distortion $d(n)$. T (or ω_0 in radians) is the period of $p(n)$ and, for the sake of simplicity, we also suppose that we have an integer number

	Mean (20-0 dB) [0 dB]		
Technique	Technique ”‘per se”’ (without oracle)	Oracle mask unvoc. and sil.	Oracle mask all
FE (Espectr.)	33.30 [7.66]	64.25 [25.04]	95.01 [90.18]
$DDR_{55,200}$ (Espectr.)	35.84 [5.84]	73.16 [37.98]	90.35 [82.75]
A. Sift ($\delta = 8$) (Espectr.)	36.61 [8.09]	77.92 [47.72]	93.36 [88.94]
N. VAD+Harm (Cocl.)	85.95 [72.21]	89.15 [73.13]	95.11 [89.40]
N. VAD+Tun (Espectr.)	87.21 [74.43]	90.87 [79.46]	95.01 [90.18]

Table 1. WAcc results for the whole Aurora-2 (Set A, B and C) obtained by four techniques which represent the four basic voiced mechanisms. 0 dB result is shown in bracket. Ideal pitch is employed.

of periods N_p ($N = N_p * T$). Its complex discrete noisy spectrum is:

$$X(\omega_k) = P(\omega_k) + D(\omega_k) \qquad (k = 0, ..., N-1) \tag{3}$$

Taking into account the periodicity of $p(n)$, the above equation can be expressed as follows:

$$X(\omega_k) = \begin{cases} P(\omega_k) + D(\omega_k) & \text{si} \quad \omega_k = \omega_0 m \\ D(\omega_k) & \text{otherwise (tunnelling samples)} \end{cases} \tag{4}$$

where $m = 0, 1, .., T-1$. From this equation, we can deduce that only a percentage $(Np-1)/Np$ of the N noise spectral samples can be recovered if we only know the pitch period T, no matter how the noisy signal is transformed. The remaining noise frequency samples are mixed with the speech harmonics and can not be recovered, although they can be estimated by applying some type of interpolation.

We can consider that the noise spectrum estimates obtained from tunnelling samples and interpolation are optimal in the sense that minimal assumptions about the noise are required (only an interpolation model). In practice, it must be also taken into account that the resulting noise estimation has some problems like non perfect periodicity or unavoidable time-window which also widens the harmonics. The reason of only taking one tunnelling sample (between the harmonics) in Pitch-based Noise Estimation technique is this widening.

4.3. Optimum voiced mechanisms

Let us consider the following three points:

1. Tunnelling noise estimate is theoretically optimum (just argued above).
2. The similarity between tunnelling and harmonicity masks (Sec. 4.1).
3. MD (with ideal mask) provides much better results than other techniques which employ a noise estimate (such as SS) (Sec. 2.3).

From these three considerations, we can say that mask estimation mechanisms based on tunnelling or harmonicity, along with MD recognition, provide a very solid framework for pitch-based recognition of voiced frames, and that in ideal conditions these can be considered as an optimum mechanisms (hypothesis H2).

5. EXPERIMENTAL RESULTS

In order to compare the robustness of the four basic mechanisms for voiced frames, WAcc (Word Accuracy) results in spectrogram (or cochleagram) domain, with ideal pitch (obtained from clean signal at each frame) and with oracle mask in unvoiced and silence frames for different techniques (representative of each mechanism) are shown in Tab. 1. The experimental framework is the typical of Aurora-2 and the ETSI FE in a spectral representation [10] (9 Gaussians per state , 23-Mel channels, etc.).

FE is used as baseline (no robust). $DDR_{55,200}$ corresponds to the asymmetric window (Sec. 2.1) and represents the mechanisms based on exploiting the harmonic structure. *A. Sift* corresponds to the sifting autocorrelation technique (Sec.

2.2) and represents the mechanisms based on comb estimation of the clean signal. *N. VAD+Harm* is the adaptation of Barker's technique (Sec. 2.4 and 3) and represents the mechanisms based on harmonicity mask estimation. *N. VAD+Tun* is the tunnelling mask (Sec. 2.3) and represents the mechanism based on tunnelling noise estimation.

The first column shows the results obtained by these techniques (all-ones mask has been employed for the first three techniques). The second column shows the same experiments but applying oracle masks to unvoiced frames and silences (this shows the success of the voiced mechanisms), and third column shows oracle mask results. The mask threshold and slope of *N. VAD+Harm* and *N. VAD+Tun* have been re-optimized to improve the results in the second column.

5.1. Confirmation of the hypothesis

The first three techniques (in the first column) could yield better results if they were implemented, not in a spectral domain, but in the cepstral domain with CMN (Cepstral Mean Normalization). In that case they could obtain the next results: 66.76 for *FE*, 77.47 for $DDR_{55,200}$ and 80.30 for *A. Sift*. Despite everything, it can be concluded that the best voiced mechanisms are the two last ones (as shown the second column), i. e. harmonicity and tunnelling mask estimations. Their results are quite similar although tunnelling is a bit better. This increment can be due to the difference between the Mel scale of the spectrogram and the ERB scale of the cochleagram. Except for this difference, it can be said that these mechanisms are similar and that they are best ones. This confirms the two previous statements made in this section (hypothesis H1 and H2).

5.2. Limits in pitch-based recognition

If we compare the first and second columns of Tab. 1 for the technique *N.VAD+Tun* and it is taken into account that second column contains an approximation to the best performance that we can obtain with the pitch-based techniques (because unvoiced and silence frames have oracle mask and voiced frames have one of the optimum voiced mechanisms) we can conclude that the proposed pitch-based noise estimation technique (first column) is almost optimum because its results are not very far from this upper boundary results (second column).

Let us compare now the second and third columns of the table. Although the results of the second column are not very far from those of the third one (oracle masks for all frames), we can see that the pitch-based mask estimation methods will never perform as well as the oracle masks (this is specially clear at 0 dB), independently of the accuracy of the pitch extractor employed.

6. CONCLUSIONS

In this paper we have presented different pitch-based techniques and classified them based on the robust mechanism applied to voiced frames. It has been shown that tunnelling and harmonicity masks are equivalence and that the mechanism of tunnelling noise achieves optimum noise estimation based on pitch and employing very little information about the noise. Taking into account this and the advantages of MD (as compared to SS) it has been pointed out the limits of the pitch-based recognition, showing that the Pitch-based Noise Estimation technique is close to these limits which yield good performance but they do not reach the oracle mask results.

This points out that in order to obtain these oracle mask results, more information than that extracted from the pitch trajectories would be required to approximate the performance of the oracle masks. This extra information could be obtained from the noise itself (dynamically updated in time from silence regions) or accurate speech models. A suitable application of this extra information to tunnelling and harmonicity mask estimation could be a future work.

7. REFERENCES

[1] J. Barker, M. Cooke, and P. Green. Robust asr based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise. In *Eurospeech*, pages 213–216, 2001.

[2] L. Buera, J. Droppo, and A. Acero. Speech enhancement using a pitch predictive mode. In *ICASSP*, 2008.

[3] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34:267–285, 2001.

[4] C. J. Darwin. Perceptual grouping of speech components differing in fundamental frequency and onset-time. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 33 (2):185–207, 1981.

[5] D. Ealey, H. Kelleher, and D. Pearce. Harmonic tunnelling: tracking non-stationary noises during speech. In *EUROSPEECH*, pages 437–440, 2001.

[6] L. Gu and K. Rose. Perceptual harmonic cepstral coefficients for speech recognition in noisy environment. In *ICASSP*, 2001.

[7] X. Huang, A. Acero, and H. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. 2001.

[8] N. Ma, P. Green, J. Barker, and A. Coy. Exploiting correlogram structure for robust speech recognition with multiple speech sources. *Speech Communication*, 49:874–891, 2007.

[9] D. Macho and Yan Ming Cheng. Snr-dependent waveform processing for improving the robustness of asr front-end. In *ICASSP*, 2001.

[10] Juan A. Morales-Cordovilla, Ning Ma, Victoria Sánchez, Jose L. Carmona, Antonio M. Peinado, and Jon Barker. A pitch based noise estimation technique for robust speech recognition with missing data. In IEEE, editor, *ICASSP (International Conference on Acoustic, Speech and Signal Processing)*, pages 4808–4811, Mayo, 22-27 2011.

[11] Juan A. Morales-Cordovilla, Antonio M. Peinado, Victoria Sánchez, and José A. Gonzalez. Feature extraction based on pitch-synchronous averaging for robust speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 19(3):640–651, Marzo 2011.

[12] Juan A. Morales-Cordovilla, Victoria Sánchez, Antonio M. Peinado, and Ángel. Gómez. On the use of asymmetric windows for robust speech recognition. *Circuits, Systems and Signal Processing (Springer)*, 2011, Abril (aceptado con cambios).

[13] Antonio M. Peinado and Jose C. Segura. *Speech Recognition over Digital Channels*. Wiley, 2006.

[14] C. Ris and S. Dupont. Assessing local noise level estimation methods: application to noise robust asr. *Speech Communication*, 34 (2):141–158, 2001.

[15] M. Seltzer, J. Droppo, and A. Acero. A harmonic-model based front end for robust speech recognition. In *EUROSPEECH*, 2003.

[16] B. Shannon and K. K. Paliwal. Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition. *Speech Communication*, 48, no. 1:1458–1485, 2006.

[17] DeLiang Wang and Guy. J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. 2006.