A PITCH BASED NOISE ESTIMATION TECHNIQUE FOR ROBUST SPEECH RECOGNITION WITH MISSING DATA

Juan A. Morales-Cordovilla[⋆], *Ning Ma*[†], *Victoria Sánchez*[⋆], *Jose L. Carmona*[⋆], *Antonio M. Peinado*[⋆], *Jon Barker*[†]

[⋆]Dept. of Teoría de la Señal Telemática y Comunicaciones, Universidad de Granada, Spain, {jamc,amp,victoria,maqueda}@ugr.es.
[†]Dept. of Computer Science, University of Sheffield, UK, {n.ma,j.barker}@dcs.shef.ac.uk.

ABSTRACT

This paper presents a noise estimation technique based on knowledge of pitch information for robust speech recognition. In the first stage the noise is estimated by means of extrapolating the noise from frames where speech is believed to be absent. These frames are detected with a proposed pitch based VAD (Voice Activity Detector). In the second stage the noise estimation is revised in voiced frames using harmonic tunnelling thechnique. The tunnelling noise estimation is used at high SNRs as an upper bound of the noise rather than a suitable estimation. A spectrogram MD (Missing Data) recognition system is chosen to evaluate the proposed noise estimation. The proposed system is compared in Aurora-2 with other similar techniques like cepstral SS (Spectral Subtraction).

***Index Terms*—** Robust speech recognition, missing data, noise estimation, VAD, harmonic tunnelling.

1. INTRODUCTION

Acoustic noise represents one of the major challenges for automatic speech recognition systems. Many different approaches have been proposed to deal with this problem [1, 2]. A powerful approach developed in the last 15 years to deal with this has been MD (Missing Data) [2, 3] and its more current extension SFD (Speech Fragment Decoding) [4, 5]. Missing data techniques adapt the conventional probabilistic ASR formalism to deal with partially corrupted data. The missing data mask – a spectro-temporal map of binary values which can be obtained by means of noise estimation – is used to label spectro-temporal pixels as being either reliable or unreliable. In this paper we will work with MD because it provides a straightforward approach to evaluate the proposed noise estimation technique in an ASR framework.

Many different approaches have been proposed for noise estimation but in essence most of them are based on a VAD (Voice Activity Detector) to detect regions of silence and extrapolate the noise to the rest of the regions [1]. One problem of this extrapolation is that with very sporadic noises it often fails. On the other hand, a technique called harmonic tunnelling estimates the background noise from voiced regions (regions with pitch) [6]. In this paper, a noise estimation technique based on extrapolation of silence regions, and revising the esimation in voiced regions by means of the tunnelling technique, is proposed. Both the proposed VAD, used to detect silence regions, and the tunnelling revision use pitch information. The structure of the paper is as follows. First, a block diagram gives an overview of the system. Section 3 explains the proposed system in greater detail and the general noise estimation scheme is also explained. The experimental framework and results on the Aurora-2 database are presented in section 4. The paper concludes with a summary and a discussion of future works.

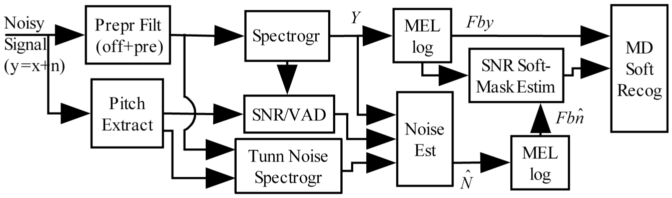

Fig. 1. Block diagram of missing data recognition system.

2. SYSTEM OVERVIEW

The recognition system (Figure 1) takes as input a noisy signal of an utterance which is the sum of clean speech and noise ($y = x + n$). *Pitch extractor* takes this signal and produces a pitch estimate for each frame. The rest of the blocks take the noisy signal passed through a preprocessing filter. This consists of an offset and a preemphasis filter which enhances high frequencies. The *SNR* (Signal to Noise Ratio estimator of the utterance) and *VAD* take as inputs magnitude spectrograms of noisy signal (Y) and the pitches. *Tunnelling noise spectrogram* estimates noise in voiced frames using the harmonic tunnelling technique [6] which makes use of the noisy signal and pitch estimates. Our center block *Noise estimator* takes Y, *SNR*, *VAD* and *tunnelling noise* to give a spectrogram noise estimation ($\hat{N}$). Y and $\hat{N}$ are passed through a MEL filter bank [7] and a log compression (which yields Fby and $Fb\hat{n}$). These two last outputs are used to estimate an SNR of each spectro-temporal pixel and consequently a corresponding soft mask. Finally, the soft mask and Fby are passed to the *MD soft recognizer* to produce a transcription of the utterance [3].

3. PITCH BASED NOISE ESTIMATION

The most important blocks and functions of the proposed system are detailed below. Note that the parameters of the blocks were determined through a preliminary experiment performed over a set of training (not testing) sentences of Aurora-2 contaminated with noise.

3.1. Extrapolation noise function

An important function which will be frequently used in this work is the extrapolation noise function defined as,

ICASSP 2011 v.. This work has been supported by the Spanish MEC/FEDER project TEC2010-18009 and partly supported by UK EPSRC grant EP/G039046/1.

$$\hat{N}_{compl}(\omega_k, t) = ExtrapN(\hat{N}_{kn}(\omega_k, t), kn(\omega_k, t)) \tag{1}$$

This function has two inputs: a noise spectrogram estimation with some known spectro-temporal pixels ($\hat{N}_{kn}(\omega_k, t)$) and a corresponding mask of this indicating with 1 that the pixel is known and with 0 that the pixel is unknown ($kn(\omega_k, t)$). The output is a complete noise estimation ($\hat{N}_{compl}(\omega_k, t)$) which mantains the same level as the input in known pixels and the value in the unknown pixels is extrapolated from known pixels. Many different extrapolation functions can be used, but for the sake of simplicity we will use the following: a current unknown pixel will be replaced by the average value of either the sequence in time of the last ten known pixels or the sequence of the next ten known pixels following a nearest neighbour criterion. The closest pixel of each last/next sequence determines the distance criterion.

3.2. Pitch extractor

Our pitch extractor is exactly the same as that employed in [8]. This pitch extractor takes the pitch provided by the ETSI xFE pitch extractor [7] and applies smoothing processing. This smoothing is needed because the pitch provided by xFE is not continuous enough (mainly at low SNRs).

3.3. Tunnelling noise spectrogram

Given a voiced frame contaminated with some background noise, it is possible to estimate the magnitude shape of the noise spectrum using samples of the noisy spectrum in the gaps between the harmonics (tunnelling samples) [6]. These samples can be used to interpolate the rest of the values in order to finally obtain a magnitude noise estimation with a desirable number of points. The magnitude spectrum of a noisy frame y with N samples is given by means of its DFT (Discrete Fourier Transform):

$$Y(\omega) = |\sum_{n=0}^{N-1} y(n)win(n)e^{-i\omega n}| \tag{2}$$

where ω indicates the frequency in radians and $win(n)$ is the window used for the spectrum estimation (in our case a hamming window). The tunnelling samples (Y(ω_l)) are obtained by evaluating equation (2) in the frequencies corresponding to the gaps. The tunnelling noise estimation, of a voiced frame t, with NFT spectral points between 0 and 2π is obtained by interpolating between these tunnelling samples:

$$\hat{N}_{tun}(\omega_k, t) = Interp(\omega_l, Y(\omega_l, t), \omega_k) \tag{3}$$

$$\omega_l = \omega_0(l + \frac{1}{2}),\ l = \{-1/2, 0, 1, 2, .., ceil(\pi/\omega_0)\}$$

$$\omega_k = \frac{2\pi k}{NFT},\ k = \{0, .., NFT/2 - 1\}$$

where ω_0 is the pitch frequency of corresponding voiced signal and $Interp$ is the interpolation function that in our case will be carried out through a linear interpolation. Figure 2 shows an example of tunnelling noise estimation. The tunnelling samples are shown with square and the tunnelling noise spectrum estimate with dash-dot line. It can be observed that tunnelling estimation is very close to real noise (dotted line). One problem of this estimation is that when the noise energy is very low compared to that of the speech signal, noise tends to be overestimated (e.g., at the two ends in Figure 2). This is because in these regions tunnelling samples follow values which are a consequence of the window used in DFT and it is impossible to recover the real noise values. This effect is not important at low SNRs but at high SNRs it is more problematic. So, at high SNRs, tunnelling noise will be used as an upper bound of the noise rather than a suitable noise estimate.

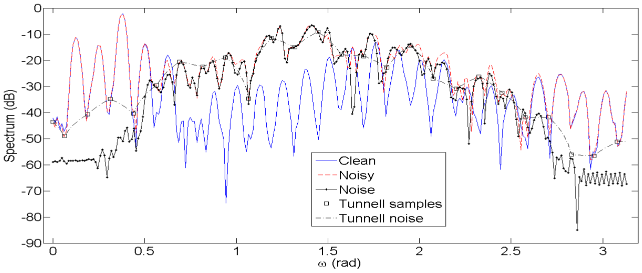

Fig. 2. Example of tunnelling noise estimation on a voiced noisy frame with pitch ω_0=0.126 rad

3.4. SNR estimator

If we have spectral estimates of the noise ($\hat{N}(\omega_k, t)$) and the clean signal ($\hat{X}(\omega_k, t)$), using Parseval's theorem, it is possible to obtain the corresponding energy ($E_{\hat{N}}(t)$ and $E_{\hat{X}}(t)$). Equation (4) shows how to estimate the SNR of the noisy utterance using these energies.

$$S\hat{N}R = 10 * log_{10}(\sum_{t \in voiced}^{nf} E_{\hat{X}}(t) / \sum_{t=1}^{nf} E_{\hat{N}}(t)) \tag{4}$$

$$\text{where } E_S(t) = \sum_{k=0}^{NFT/2-1} |S(\omega_k, t)|^2 \tag{5}$$

where nf is the number of frames. Only voiced frames are used to estimate the total energy of the clean signal. We do this because to estimate the energy of the clean signal, in Aurora-2, no silence regions are taken into account (*ITU* recommendation P.56) and using only voiced frames gives very similar results because voiced frames contain the most part of speech energy. In order to obtain $\hat{N}(\omega_k, t)$ we assume that speech is absent in the first and last ten frames of the noisy spectrogram ($Y(\omega_k, t)$). These two known noise regions are passed to the extrapolation function (1) to obtain a complete estimation of the noise spectrogram. The clean spectrogram is estimated through a simple spectral subtraction: $\hat{X}(\omega_k, t) = Y(\omega_k, t) - \hat{N}(\omega_k, t)$ (0.06 is taken as floor value).

3.5. Voice Activity Detection

Our VAD is based on the previous pitch extractor and detects three different classes of frames: silence, unvoiced and voiced. Frames labeled as voiced correspond to frames where the pitch extractor gives a valid pitch. Unvoiced frames correspond to speech sounds like fricative, plosive, etc., and we assume that they have two properties [9]: first, their energies are mainly between 1800 and 4000 Hz and, second they can be only found after or before a sequence of voiced frames and never occur in isolation. Following the first property, an instantaneous SNR of high frequencies (HF) can be estimated as,

$$S\hat{N}R^{HF}(t) = 10 * log_{10}(E_{\hat{X}}^{HF}(t) / E_{\hat{N}}^{HF}(t)) \tag{6}$$

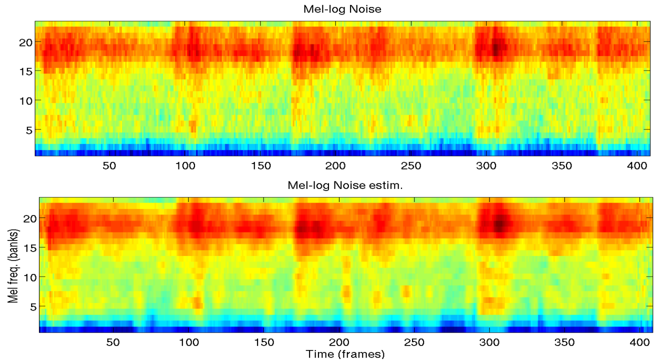

Fig. 3. Subway Mel-log noise and its estimation of the Aurora-2 utterance 4460806 at 0dB.

where the clean $\hat{X}$ and noise $\hat{N}$ spectrograms are estimated in the same way as in section 3.4, by means of a simple spectral subtraction. The frame energies are estimated using equation (5) but instead of summing over the full range of frequencies, only the frequencies between 1800 and 4000 Hz are employed. Taking the second aforementioned property into account and this instantaneous SNR measure, we consider the frames with $S\hat{N}R^{HF}(t) > 3dB$ and which occur just 20 frames after or before a sequence of voiced frames as unvoiced frames. Subsequent experiments have also shown that at low SNRs, this unvoiced estimation takes many noise frames as unvoiced. So when $S\hat{N}R < 10dB$, it is assumed that unvoiced signals are mixed with noise and no detection of unvoiced frames is carried out. Finally, silence frames are those that have been classified neither as voiced nor unvoiced.

3.6. Noise estimator

Our noise estimation is perfomed in two stages. In the first stage, it is supposed that in silent regions (detected with our VAD) the noisy spectrogram ($Y(\omega_k, t)$) is dominated by noise, so that these known regions of noise are passed to the extrapolation function (1) to obtain a first estimation of the noise. In the second stage the corresponding voiced frames of the first estimation are revised using our tunnelling noise estimation. As mentioned in section 3.3, tunnelling noise provides a good estimation of noise when the SNR is low but at high SNRs, it is better to use tunnelling noise as an upper bound of the real noise. Following this idea, when $S\hat{N}R < 10dB$ the voiced frames of the first noise estimation are replaced by tunnelling noise, otherwise tunnelling noise is used as an upper bound for these frames. Finally, the noise revised in the second stage is passed through a temporal mean filter of 5 frame long to smooth possible errors and the final product is our proposed spectrogram noise estimation. Figure 3 shows a comparative example of this estimation.

3.7. SNR Soft Mask Estimator

In order to obtain a soft mask for every Mel-log noisy pixel ($Fby(m, t)$), the SNR of every pixel is obtained by,

$$S\hat{N}R(m,t) = 20 * log_{10}(e^{Fb\hat{x}(m,t)}/e^{Fb\hat{n}(m,t)}) \qquad (7)$$

where $Fb\hat{n}(m, t)$ is our noise estimate (in the mel filterbank log-output domain) and where the clean $Fb\hat{n}(m, t)$ is estimated by means of a simple spectral subtraction after undoing the log compression: $e^{Fb\hat{x}(m,t)} = e^{Fby(m,t)} - e^{Fb\hat{n}(m,t)}$ where 0.06 is taken as the floor value. The soft mask is generated by compressing $S\hat{N}R(m, t)$ between $[0, 1]$ with a sigmoid function [3]. The values of threshold and slope of this function are $\beta = -3$ (i.e. SNR -3 dB) and $\alpha = 0.2$ respectively, and they have been determined empirically.

4. EXPERIMENTAL FRAMEWORK AND RESULTS

4.1. Experimental framework

The experiments reported here employ the Aurora-2 speaker independent connected digit recognition task. A very similar parametrization to that employed in [7, 10] is used and can be summarized as follows: sampling frequency 8000 Hz, frame shift and length 10 and 32 ms, 512 spectral points (range [0,2π]) and 23 Mel channels. Dynamic features are also employed, obtaning a final feature vector with 46 dimensions. Speaker independent models are trained using the clean training set. The Aurora model topology and training regime has been adhered to, but instead of using 3-component mixtures (typical of cepstrum parametrization) 9-components mixtures are used because the features employed here do not have the near independent of the cepstral features. Marginalization-based soft missing data decoding [2] is employed in the recognition engine.

4.2. Experimental results

Table 1 shows the different word accuracies achived by different systems tested over the whole Aurora-2 database.

The first three systems, labeled with *Ceps*, employ as input to the recognizer a cepstral vector estimation of clean signal. Each vector consists of 39 coefficients (13 statics, 13 deltas and 13 deltas-deltas). They are extracted according to ETSI front-end standard [7]. In addition, CMN (Cepstral Mean Normalization) is also applied. *FE* is the standard feature extractor [7] which applies no robust mechanism to estimate clean signal, only CMN. *Sift* is presented here as an example of cepstral front-end which employs the pitch to perform robust speech recognition. It estimates the autocorrelation of clean signal by means of the so-called sifting technique which uses pitch information and the supposition that the noise is uncorrelated [8]. *N. prop., SS* is when the proposed noise estimation is used in an SNR-dependent SS (Spectral Subtraction) to estimate clean signals. This uses an attenuation factor of $A = 10dB$. SNR estimation and H_{ss} filter are smoothed over time and frequency respectively with a mean filters of 9 frame long to reduce musical noise distorsion [11].

The last four systems, labeled with *MD*, follow the above explained MD configuration (section 2). They employ as input to the recognizer the Mel-log noisy spectrum and its corresponding soft mask. *N. first-last 10 frames (MD)* is when the first and last ten frames are used to estimate the noise as in section 3.4. *N. prop., no tun.* is when the proposed noise estimation uses only silent regions from the VAD to estimate it, i.e. when tunnelling estimation is not used and only the first stage of the proposed noise estimation is applied (section 3.6). *N. prop.* is the proposed noise estimation and *N. prop., a priori pitch* is the same but with pitch always obtained from the corresponding clean signal.

It can be observed that the proposed noise estimation, in MD configuration, gives the best average result (81.55). The same noise estimation but with cepstral features (SS) provides worse results (74.35) mainly at very low SNRs. It is due to the simplicity of the conventional SS scheme applied. This scheme needs the assumption that the power spectrum of the noise was to be, in general, at a lower

Systems	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Mean (20-0 dB)
FE (Ceps)	99.12	97.17	92.53	76.15	44.16	23.02	13.00	66.61
Sift (Ceps)	98.80	96.63	94.45	89.36	76.39	45.37	14.84	80.44
N. prop., SS, (Ceps)	99.36	96.66	92.09	81.84	64.09	37.06	9.72	74.35
N. first-last 10 frames (MD)	98.73	94.95	90.08	80.72	65.04	43.37	21.85	74.83
N. prop., no tun. (MD)	98.75	96.05	93.24	88.17	77.96	50.90	13.36	81.26
N. prop. (MD)	98.76	95.82	92.10	86.80	78.52	54.53	16.14	81.55
N. prop., *a priori* pitch (MD)	98.76	95.86	92.70	88.68	84.36	74.83	56.45	87.29

Table 1. Word accuracies obtained by different systems tested with Aurora-2 (Set A, B and C) for different SNR values.

level in magnitude than that of the speech, and this is not satisfied at low SNRs. A more efficient SS scheme specifically tuned to low SNR as in [12] could have been employed to solve this problem, but MD with a simple SNR approach also gives good results and this is the reason of its selection.

If *N. prop., no tun.* is compared with *N. prop.* it can be thought that the addition of tunnelling noise does not provide a great benefit (mainly at high SNRs). It is mainly because Aurora-2 noises are known of being on a whole, quite stationary, but in more sporadic kind of noises this addition could potentially provide greater benefit. This can be observed across different types of noise. E.g. in car noise, which is more stationary than speech babble, exploiting the harmonic tunnelling based noise estimation improves the accuracy at 0 dB only 2.7 percents compared to 5.6 percents in babble. Finally, *N. prop.* does not achieve as good results as *N. first-last 10 frames (MD)* at -5 dB largely due to errors in pitch estimation. This problem could be solved with a better pitch extractor as the result with *a priori* pitch shows.

5. CONCLUSIONS

In this paper a pitch based noise estimation has been presented. This estimation is first carried out by means of extrapolating the noise from silence frames. Then, the first estimation is revised in voiced frames using tunnelling noise estimation. In order to detect silence frames, a VAD has been proposed. The VAD uses the pitch frames in order to identify unvoiced frames around them. The frames which are classified as neither voiced nor unvoiced are considered as silence with noise present. Tunnelling revision is employed to make the estimation more robust in the case of very sporadic noise. Tunnelling noise is used as an upper bound at high SNRs and as a suitable noise estimation at low SNRs. Finally, the performance of the proposed noise estimation has been evaluated in Aurora-2 over a cepstral SS and a spectrogram MD recognition system. The MD system has provided better results mainly at low SNRs.

Regarding future work, the results at high SNRs could be improved specially in the detection of unvoiced frames because the proposed VAD is very simple and at these levels these kind of frames are more important. The results with *a priori* pitch show that the improvement of the pitch detector at low SNRs is an important issue. The pitch based noise estimation technique assumes the speech is the only harmonic source present – this assumption is often not true in a real situation. One solution is to consider several pitch candidates at each frame, and each candidate could result in a different noise estimation hypothesis. These parallel hypotheses can be evaluated separately using the missing data technique employing the mask derived from a hypothesized noise estimate and the one that gives the greatest likelihood is chosen. This is similar to the speech fragment decoding idea [4] and uses top-down speech models to resolve bottom-up signal ambiguity.

6. REFERENCES

[1] A. M. Peinado and J. C. Segura, *Speech Recognition over Digital Channels*, Wiley, 2006.

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

[3] 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 *Proc. Eurospeech*, 2001, pp. 213–216.

[4] J. Barker, M. Cooke, and D. Ellis, “Decoding speech in the presence of other sources,” *Speech Communication*, vol. 45, pp. 5–25, 2005.

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

[6] D. Ealey, H. Kelleher, and D. Pearce, “Harmonic tunnelling: tracking non-stationary noises during speech,” in *Proc. Eurospeech*, 2001, pp. 437–440.

[7] ETSI ES 202 211 v1.1.1., *Distributed speech recognition; Extended front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm*, July 2001.

[8] J. A. Morales-Cordovilla, A. M. Peinado, V. Sanchez, and J. A. Gonzalez, “Feature extraction based on pitch-synchronous averaging for robust speech recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, 2011.

[9] J. Ryalls, *A basic introduction to speech perception*, Speech Science Series, 1997.

[10] D. Pearce and H.-G. Hirsch, “The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *Proc. ICSLP*, 2000, vol. 4, pp. 29–32.

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

[12] J. Beh and H. Ko, “A novel spectral subtraction scheme for robust speech recognition: spectral subtraction using spectral harmonics of speech,” in *Proc. IEEE ICASSP*, 2003, vol. 1, pp. 648–651.