ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

USE OF VOICING INFORMATION TO IMPROVE THE ROBUSTNESS OF THE SPECTRAL PARAMETER SET*

Dušan Macho and Climent Nadeu

TALP Research Center
Universitat Politècnica de Catalunya, Barcelona, Spain
Email: {dusan, climent}@talp.upc.es

ABSTRACT

Speech recognition systems that operate in real world environments have to be robust against additive noises. In this work, a technique that uses a voicing-dependent exponent in the computation of the filter-bank parameters to improve their robustness to additive noise is presented. Speech recognition experiments with the Aurora 1.0 database and recognition setup are reported to show the potential of this new technique.

1. INTRODUCTION

The use of speech recognition systems in real-world operating environments requires techniques that are robust in front of additive noises. Many techniques have been proposed for that purpose [1]. Recently, the voiced/unvoiced (V/U) information was used in noisy speech recognition to apply different speech enhancement techniques to voiced and unvoiced frames [2].

In general, the effect of additive noise on the speech spectrum is more remarkable at frequency bands or time segments where the speech spectrum shows low amplitude. The speech low-power spectral regions that will be considered in this paper are the between-harmonic valleys of voiced sounds and the whole band of unvoiced sounds. As silence segments have low power in absence of noise, in this work they will be grouped with unvoiced frames and will not be distinguished from them.

If the spectrum is expressed in dB, the added noise increases the spectral values at those low-power spectral regions or time segments more than it does with the high-power ones, so the amplitude contrast of the spectrum decreases along frequency and also along its time evolution. In this paper, the voicing information is used to restore that amplitude contrast up to some extent. The character of each frame is introduced explicitly in the computation of the filter-bank (FB) parameters by means of an exponent that depends on it.

In Section 2, we show by simulation that assuming an harmonic spectral structure and additive white noise, there exists a value of the exponent involved in the computation of the FB parameter for which the mismatch on the log FB parameter is minimum. Two different ways of applying that voicing-dependent exponentiation are presented in Section 3, and recognition results with the Aurora 1.0 database are reported in Section 4.

2. VOICING-DEPENDENT EFFECT OF THE EXPONENTIATION ON THE LOG FB PARAMETER MISMATCH

Generally, the k^{th} subband log FB parameter $S(k)$ is obtained in the following way,

$$S(k) = \log\left[\sum_i W_k(i)\left|X(i)\right|^\gamma\right] \tag{1}$$

where $X(i)$ is the value of the i^{th} FFT bin and $W_k(i)$ is a mel-scaled frequency weighting function that defines the subband where the FFT magnitude values are integrated. The exponent γ should be equal to 2 in order to obtain actual energy values, but it is usually set to 1 for clean speech [3], although a value 2 seems preferable for noisy speech provided that the noise is not speech-like (Cf. Table 1).

Let us investigate now whether there exists a value of γ that is optimum in terms of the mismatch between the log FB outputs of a noisy speech frame and those of the corresponding clean speech frame according to a given reasonable measure. Assuming the harmonic spectral structure of a voiced frame that is contaminated by additive white noise within a FB subband, a simulation experiment was performed to ascertain whether there exists a value γ inside of the interval $0.1 \le \gamma \le 5.0$ for which the difference (*mismatch*) between the noisy log FB output $S_N(\gamma)$ and the clean log FB output $S(\gamma)$ expressed as

$$\begin{gathered} D(\gamma) = S_N(\gamma) - S(\gamma) = \\ \log\left[\sum_i \left|X(i) + N(i)\right|^\gamma\right] - \log\left[\sum_i \left|X(i)\right|^\gamma\right] \end{gathered} \tag{2}$$

is minimal. Note that in equation (2) we simplified the expression from (1) by dropping the subband index k and assuming a rectangular frequency window.

In our simulation, the signal consisted of a train of impulses corrupted by additive white noise. As we are simulating the computation of $S(k)$ at one subband, to use that signal is equivalent to assume a flat spectral envelope in that subband. Figures 1(a) and 1(b) show the FFT magnitude and power spectra, respectively, of that synthetic frame within the considered subband (its width has not any relevance for this simulation).

We can observe that, when $S(k)$ is computed from the power spectrum, i.e. γ=2, there is a higher contrast between harmonic

* This work has been supported by CICYT, under projects TIC98-0683 and TIC98-0423-C06-01.

peaks and between-harmonic valleys than in the case of using magnitude spectrum, i.e. γ=1, so the noisy valleys will relatively contribute less to the FB summation. However, the harmonic peaks, whose contribution has been increased, become noisier. It can be expected that this tradeoff will be reflected in the difference $D(\gamma)$ so there will be an optimum γ that minimizes it.

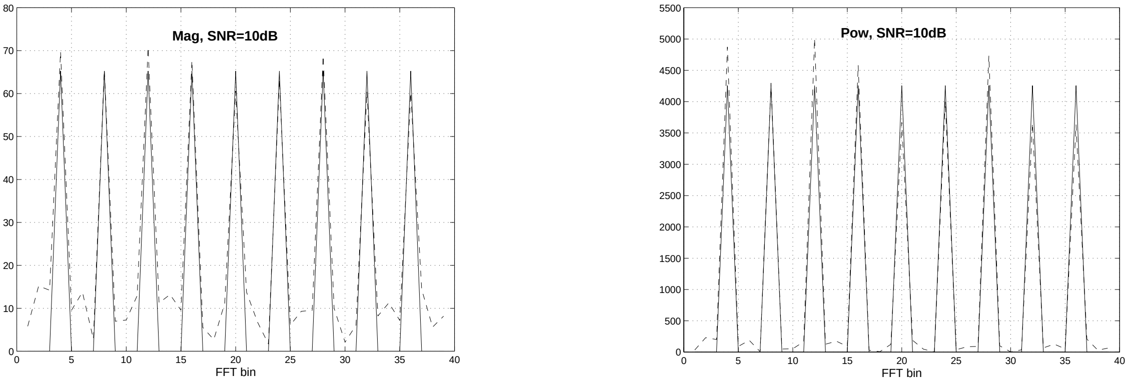
The difference $D(\gamma)$ from (2) was computed for several SNRs and for two fundamental frequencies, and it was averaged over 100 realizations of the simulated signal. Figures 2(a) and 2(b) show the average difference as a function of γ for each one of the fundamental frequencies and SNRs. Notice that each curve has a minimum along γ. The position of the minima depends on both SNR and F_0 (and also on the envelope shape within the subband) but the curves suggest that a γ exponent larger than 1 should probably be used for noisy voiced frames to reduce the mismatch. Additionally, the optimal values of γ are higher for a larger F_0.

The minima in the $D(\gamma)$ function are caused by the harmonic peak structure and, therefore, they would not arise for simulated unvoiced frames. Actually, assuming a flat subband spectrum, the use of a larger γ increases the $D(\gamma)$ difference. These observations led us to think that if a value of γ larger than 1 had been found useful for noisy speech recognition (Cf. Table 1) it was because of the voiced frames, so the recognition performance would benefit from using a value of γ that depends on a voiced/unvoiced classification; in particular, a γ larger for voiced frames than for unvoiced frames.

3. USE OF VOICING-DEPENDENT FB PARAMETER COMPUTATIONS FOR RECOGNITION OF NOISY SPEECH

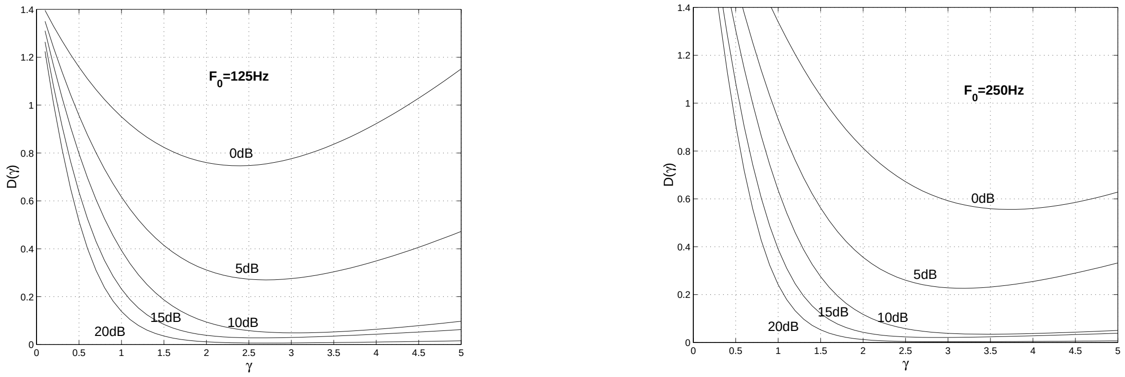
In the above section, we have seen how the use of a γ exponent that depends on the V/U decision may help to reduce the mismatch between both clean and noisy speech representations. Although that mismatch reduction might not imply an increment of the speech recognition performance, since the capacity of the recognizer for discriminating speech sounds may be debilitated by that operation, digit recognition experiments reported in Section 4 will show an improvement in recognition performance.

However, in the simulation of the above section we have only taken into account the computation of a FB parameter in a given band, but not the temporal sequence of that parameter. As the γ exponent should be larger for voiced frames than for unvoiced frames, it actually increases the amplitude contrast between voiced and unvoiced frames in the time sequences of log FB parameters, which results in an increased separation between voiced and unvoiced regions in the parameter space. This effect is illustrated in Figure 3(a), where the time evolution of a log FB parameter amplitude is depicted. Whereas additive noise tends to



(a) (b)

Figure 1 (a) Magnitude spectrum (γ =1), and (b) power spectrum (γ=2) of clean (solid line) and noisy (dashed line) simulated voiced frame. Assuming an 8 kHz sampling rate, the plotted first 40 FFT bins correspond to the band 0-1250 Hz. A flat spectral envelope in that band and a fundamental frequency F_0=125 Hz were assumed.



(a) (b)

Figure 2 Difference between noisy and clean log FB parameters (see equation (2)) as a function of the exponent γ. Two fundamental frequencies are considered: (a) F_0=125 Hz (male), and (b) F_0=250 Hz (female).

decrease the original amplitude gap between voiced and unvoiced/silence portions of speech parameters (Fig. 3(a)), the V/U contrasting operation restores it to some extent in relative terms (Fig. 3(b)).

As this time domain contrasting effect may also be responsible for the recognition performance improvement, we have considered an alternative way of obtaining it in order to ascertain whether the performance improvement is due either to the reduction of mismatch in the computation of the FB parameters or to the increment of the amplitude contrast of the time sequences of FB parameters. This alternative way consists of applying an exponentiation after the FB computation and before the log compression. Note that this operation is equivalent to a multiplication after the log and it is applied to a reduced number of parameters. In this case, there is not a different behavior between voiced and unvoiced frames in terms of frequency mismatch. In order to distinguish both exponentiations we will herewith use the notation γ^{FFT} for the exponentiation applied on the FFT magnitude values and γ^{FB} for the exponentiation applied on the FB parameters. Recognition results are presented in the next section for both types of exponentiations.

4. EXPERIMENTAL EVALUATION

4.1 Database, Recognition System and V/U Detection

The Aurora 1.0 database, that is being used for developing the ETSI noisy speech Distributed Speech Recognition (DSR) standard front-end [4], was employed in this work for testing the described exponentiation techniques. This corpus consists of the TI connected digit utterances, downsampled from 20 kHz to 8 kHz, and with artificially added noises at the following SNR levels: clean, 20, 15, 10, 5, 0, and –5 dB. Four types of real noises have been used: hall, babble, suburban train, and car.

The recognition system, which is the one used for the above standardization work, is based on continuous density HMMs with diagonal covariance matrices. Each digit is modeled by 18 states with 3 Gaussians per state and silence is modeled by 6 states with 6 Gaussians per state. In the front-end, the parameters are computed from 30 ms Hamming-windowed frames with 10 ms shift. 14 log FB magnitude/energy parameters are computed from each frame by integrating the FFT magnitude or power spectra according to mel-scaled triangular frequency windows with 50% overlapping. Then, frequency filtered (FF) parameters [5] are computed for each frame by using a second order frequency filter z-z^{-1}. In the tests with only static features, we do not use the two endpoints of the FF parameter vector that actually are absolute log FB magnitudes/energies of the 2nd and 13th FB bands, to avoid their influence on the recognition performance.

In the tests with voicing-dependent exponentiation, we use a very simple V/U detection based on the spectral slope at each frame, which is estimated by approximating the spectrum with a first order polynomial and using the mean-square error criterion. The V/U detection is always performed on the clean version of each utterance, and the use of a slope threshold has shown to be sufficient for this purpose. Consequently, the presented results may be considered as scores that could be obtained with an error free V/U detection on noisy speech.

We tested the voicing-dependent exponentiation techniques for two training modes: clean speech training (CST) and multicondition training (MCT). In CST, only clean digit utterances have been used for training. In MCT, digit utterances corrupted by all four noises at SNRs equal to 20, 15, 10, and 5 dB have been employed in training along with the clean ones.

4.2 Tests with Static Features

In all the following tables, the clean and average noisy (over SNRs equal to 20, 15, 10, 5, and 0 dB) word accuracy rates are reported. Also, the total average word accuracy for noisy speech is shown in the last column of each table.

Table 1 shows the baseline performances for both CST and MCT modes. Two different sets of results are shown for each training mode: one by using the magnitude FFT spectrum (Mag) and other by using the power FFT spectrum (Pow). We can observe that, in CST, an improvement in recognition performance can be obtained by using the power spectrum instead of the magnitude spectrum in all noisy conditions, except for babble noise. As the babble noise possesses a harmonic structure, it may be emphasized by using a higher γ.

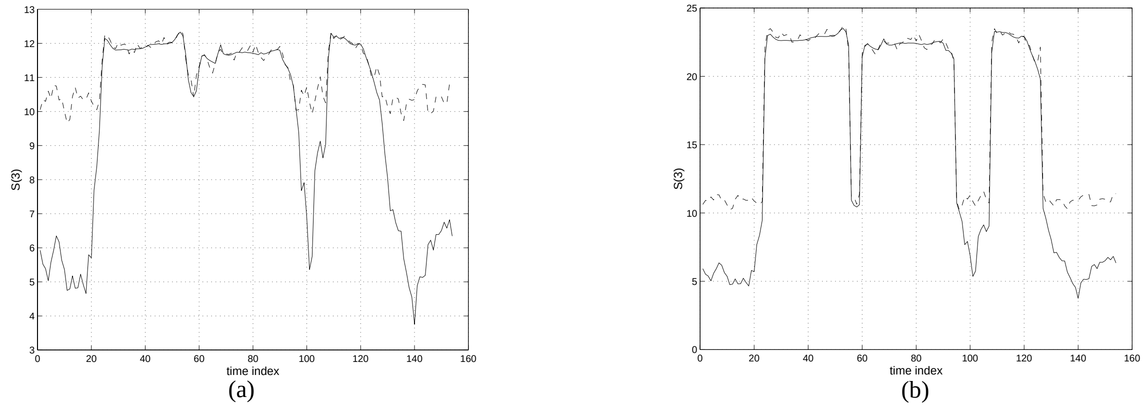


Figure 3 Time sequence of a log FB parameter (3rd subband) extracted from a 14-order FB analysis. (a) log FB amplitude with γ=1, and (b) same log FB parameter with V/U dependent γ^{FFT}, where γ^{FFT}=1 was used for unvoiced frames and γ^{FFT}=2 for voiced frames. Solid lines represent the clean speech parameter sequence and dashed lines represent the parameter sequence obtained from speech corrupted by additive hall noise for SNR=10 dB. The speech signal corresponds to a male utterance containing the digit sequence "five nine four".

In MCT, the noisy speech recognition rates for both magnitude and power are more similar than in CST. The largest improvement obtained by the power corresponds to the car noise case. As expected, MCT yields a considerable improvement in all noisy conditions with respect to CST. However, for clean speech recognition its performance is lower than CST, due to the mismatch between training and testing conditions.

	Clean	Hall	Babble	Train	Car	Noisy Aver.
CST, Mag	95.14	46.36	40.71	46.66	65.75	49.87
CST, Pow	95.02	48.45	39.10	50.36	67.65	51.39
MCT, Mag	92.02	60.26	56.92	71.13	72.21	65.13
MCT, Pow	91.50	59.88	56.73	72.28	75.09	66.00

Table 1 Baseline CST and MCT digit recognition percentages with the Aurora 1.0 database, using either spectral magnitude (γ=1) or power (γ=2).

By performing a few tests with voicing-dependent γ, we found a good performance using $\gamma^{FFT}=\gamma^{FB}=1$ for unvoiced frames and $\gamma^{FFT}=\gamma^{FB}=2$ for voiced frames. Comparing Tables 1 and 2, we can observe that, for both CST and MCT training modes, the average noisy speech recognition rate have been considerably improved by using voicing-dependent exponentiation. Using γ^{FFT} instead of magnitude, it has increased, for CST, from 49.87% to 64.29%, and, for MCT, from 65.13% to 80.48%. The highest improvement is achieved for babble noise; however, note that the V/U decision was –artificially– made with clean speech so, in that case, the V/U-dependent exponentiation can reduce the mismatch in unvoiced segments of the target speech signal that have been contaminated by voiced noise.

	Clean	Hall	Babble	Train	Car	Noisy Aver.
CST, γ^{FFT}	95.61	58.48	61.87	58.63	78.19	64.29
CST, γ^{FB}	95.69	56.38	60.52	57.12	75.88	62.48
MCT, γ^{FFT}	94.02	73.17	79.87	82.35	86.54	80.48
MCT, γ^{FB}	94.16	73.11	80.16	82.97	85.13	80.34

Table 2 CST and MCT digit recognition percentages with the Aurora 1.0 database, using either FFT or FB voicing-dependent exponentiation.

Note that, in matched conditions, i.e. for CST clean speech recognition, the scores have also been slightly improved by using a voicing-dependent γ. When MCT is employed, the clean digit recognition rates improve noticeably by using any of the two voicing-dependent exponentiations.

For unmatched conditions, there is not a large difference between the two types of variable exponentiation, although an advantage for γ^{FFT} in front of γ^{FB} can be observed for CST. Consequently, from these results it appears that the predominant effect is the accentuation of contrast between voiced and unvoiced/silence segments in the temporal sequences of spectral parameters.

4.3 Test with both Static and Dynamic Features

In preliminary experiments with both static and dynamic parameters, worse recognition results were found when using the voicing-dependent exponentiation. In particular, the clean speech recognition rate decreases considerably when computing the dynamic features from the V/U-enhanced-contrast log FB parameters. Thus, we used the V/U-dependent γ only in the static parameter set while we eliminated it from the dynamic feature sets. In order to avoid the computation of a separate static log FB parameter set for computing dynamic features, we just performed a de-exponentiation on each V/U-enhanced-contrast frame. This procedure is straightforward in the case of γ^{FB} since it consists of a multiplication. In the case of γ^{FFT} we use the same strategy, so that, as the de-exponentiation is applied after the FB summation, it is not an exact inverse operation.

	Clean	Hall	Babble	Train	Car	Noisy Aver.
MCT, Mag	98.89	86.70	83.60	87.10	93.45	87.71
MCT, Pow	98.99	86.00	82.80	88.35	93.92	87.77
MCT, γ^{FFT}	98.51	87.00	89.37	91.08	94.75	90.55
MCT, γ^{FB}	98.58	87.71	90.09	90.87	94.79	90.86

Table 3 MCT baseline results (the first two rows) and results obtained by the two types of voicing-dependent exponentiation (the last two rows) when derivative and acceleration features are included.

Only MCT tests were carried out in this case. 14 FF static parameters with both absolute energy endpoints were included. Two Slepian time filters [6] with parameters K_1=1, L_1=5, W_1=20, and K_2=2, L_2=6, W_2=20 were employed, in combination with the equalizer 1-0.97z^{-1}, for computing the first and second dynamic parameter sets (in total, 42 parameters per frame). We can observe from Table 3 that an improvement can be obtained in all noisy conditions by using either γ^{FFT} or γ^{FB}.

5. CONCLUSIONS

Using a V/U-dependent exponent for computing the FB parameters, a significant recognition performance improvement is obtained with static features for the Aurora database and recognition setup. After comparing the results obtained from two different exponentiation techniques, it is concluded that the predominant effect is the enhancement of the amplitude contrast between voiced and unvoiced/silence segments in the temporal sequences of spectral parameters.

6. REFERENCES

[1] J.-C. Junqua, J.-P. Haton, *Robustness in Automatic Speech Recognition*, Kluwer, 1996.

[2] D. O'Shaughnessy, H. Tolba, "Towards a Robust/Fast Continuous Speech Recognition System Using a Voiced-Unvoiced Decision", Proc. ICASSP, 1999.

[3] ETSI SQL WI007, http://webapp.etsi.org/WorkProgram/Report_WorkItem.asp?WKI_ID=6400.

[4] D. Pearce, "Experimental Framework for the Performance Evaluation of Distributed Speech Recognition Front-ends", Aurora Document Number AU/120/98, Sept. 1998.

[5] C. Nadeu, D. Macho, J. Hernando, "Time & Frequency Filtering of Filter-Bank Energies for Robust HMM Speech Recognition", to appear in Speech Communication, 2000.

[6] C. Nadeu. P. Paches-Leal, B.-H. Juang, "Filtering the Time Sequence of Spectral Parameters for Speech Recognition", Speech Communication 22, 1997, pp. 315-322.