INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

More robust J-RASTA processing using spectral subtraction and harmonic sieving

Hiroaki Ogawa[*][†]
hiro@icsi.berkeley.edu

TR-97-031

August 1997

Abstract

We investigated spectral subtraction (SS) and harmonic sieving (HS) techniques as preprocessing for J-RASTA processing to achieve more robust feature extraction for automatic speech recognition. We confirmed that spectral subtraction improved J-RASTA processing, and showed that harmonic sieving additively improved J-RASTA+SS. We investigated the performance with the Bellcore isolated digits task corrupted with car noise (additive noise) and linear distortion filter (convolutional noise). The J-RASTA+SS+HS system reduces the word error rate by 39% given pitch estimated from clean speech, and 35% given pitch estimated from corrupted speech. The system was also tested with several kind of noises from the NOISEX92 database; each noise sample was added with speech for a resulting of 0dB signal to noise ratio. SS significantly reduced word error rate for all type of noises (white noise 39%, pink noise 51%, car noise 78%, tank noise 59%, and machine gun noise 19%). Given correct pitch, HS additively reduced the word error rate for the first three noises (white noise 7%, pink noise 16%, and car noise 17%).

[*] Sony Corporation D21 Laboratory, 6-7-35 Kita-Shinagawa, Shinagawa-ku, Tokyo, JAPAN

[†] On visiting EE Division, Department of EECS, University of California, Berkeley

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Robustness of automatic speech recognition (ASR) systems under adverse environmental conditions is crucial for real applications, and various approaches have been proposed. One of the major schemes is noise modeling. The modeled noise is generally used to compensate the input signal, or to adapt the recognizer itself to the noisy environment. One well known parametric noise modeling for ASR is the parallel model combination (PMC) technique [1, 2]. Given a sample of background noise for training the model, it works very well with an HMM system. A non-parametric noise compensation, spectral subtraction (SS), is widely used in many speech applications. Although it requires estimation of the background noise and ad hoc spectral operations such as flooring, it has two merits: fewer assumption about the noise than a parametric model, and potential use as a preprocessing step to other methods. In fact, SS improves PMC+HMM systems in highly noisy environments [3].

An enhancement of dynamic features of the speech signal is another approach for robust feature extraction. The enhancement consequentially suppresses quai-stational background noise. Cepstral mean subtraction (CMS) and RASTA processing [4] are well known as robust feature extraction techniques that enhance dynamic features of speech. These techniques effectively remove the spectral distortion (i.e., convolutional noise), which typically are introduced by the transmission channel (microphones, telephone lines, etc..). J-RASTA [1] [5] processing addresses the convolutional noise and additive noise. J-RASTA balances the effect of convolutional noise with that of additive noise, and reduces both of them based on the signal to noise ratio (SNR). J-RASTA is very robust in some adverse environments [4]. However, since it is a trade-off between two kind of effects, it can not compensate for both of them completely. In other words, it could improve the robustness of J-RASTA to compensate the spectrum before J-RASTA in linear or logarithmic domains.

Separation of speech from noisy signals based on typical structures of speech such as harmonic structure have been investigated, especially in speech enhancement [6, 7, 8, 9]. It is another way to compensate the corrupted speech. The possibility of improving ASR systems has been suggested, but has not been evaluated with ASR very much. We apply a harmonic sieving technique (HS) and a spectral subtraction (SS) technique as preprocessing of J-RASTA feature extraction. This improves robustness of J-RASTA processing, especially for speech with both additive & convolutional or with high additive noise.

[1] J-RASTA is sometimes referred to as "LinLog-RASTA"

Chapter 2

Methods

2.1 J-RASTA processing

In this section we briefly review J-RASTA processing. We assume the observation signal $y(t)$ as follows:

$$y(t) = h(t) * (x(t) + d(t))$$

where $x(t)$ is a pure speech signal, $d(t)$ is additive noise, and $h(t)$ is convolutional noise. In the logarithmic magnitude spectral domain, we can linearly separate the convolutional noise as follows:

$$\log Y(\omega) = \log H(\omega) + \log(X(\omega) + D(\omega)) \tag{2.1}$$

Since power spectrum of convolutional noise in the real world such as communication channel distortion is changes relatively slowly in comparison to speech, we can remove the first term "$\log H(\omega)$" in equation (2.1) by high pass filtering. In Log-RASTA processing, the following band-pass filter is applied to the logarithmic magnitude spectrum (i.e., $\log Y(\omega)$). This band pass filter also suppresses the rapid spectral change which is unseen in speech signal.

$$R(z) = 0.1z^4 * \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}}$$

Then, we can extract a term "$\log(X(\omega) + D(\omega))$", which is affected by additive noise $D(\omega)$.

In the power spectral domain, we can linearly separate speech which is only affected by convolutional noise.

$$Y(\omega) = H(\omega)X(\omega) + H(\omega)D(\omega) \tag{2.2}$$

By using the same scheme of Log-RASTA processing, we can reduce the additive noise (i.e., $H(\omega)D(\omega)$) which changes relatively slowly or rapidly, but the output is affected by convolutional noise and additive noise which comes through the filter $R(\omega)$.

Two issues come up in this case. One is that we need to reduce both additive and convolutional noise in the same time. The other is the noise which comes through $R(\omega)$ in equation (2.2).

J-RASTA processing is a solution of the former problem. It balances the effects of equation (2.1) and equation (2.2) by mapping the input spectrum as follows:

$$\hat{X}(\omega) = \ln(1 + JX(\omega))$$

J is set to be close to zero if additive noise is dominant, otherwise J should be a larger value to suppress convolutional noise. In other words, convolutional noise is more reduced when when J is set to larger value, because in this case $X'(\omega)$ is log-like.

Though J-RASTA works well for both additive and convolutional noise, the scheme of J-RASTA is a trade-off of reduction of errors due to convolutional noise and additive noise. It is still difficult for J-RASTA to suppress both of them in the case that both convolutional and additive noise have large effects. Thus, it is expected that spectral compensation in the linear domain or in the logarithmic domain before J-RASTA processing could improve the robustness, especially in very adverse environments.

2.2 Spectral Subtraction (SS)

Spectral subtraction (SS) is widely used for additive noise suppression [10, 3, 11]. Because of its simplicity, SS is easy to use for noise suppression and it can work well as frontend processing with another feature extraction technique. The parallel model combination technique, which has been shown to make HMM ASR systems more robust, is also improved by spectral subtraction in highly noisy environments [3]. Similarly, we use the SS as preprocessing for J-RASTA in order to improve the robustness.

The scheme of spectral subtraction we use is defined as follows

$$Y_{ss}(\omega) = \max(Y(\omega) - \alpha\tilde{N}(\omega), \beta Y(\omega)) \tag{2.3}$$

where $\tilde{N}(\omega)$ is estimated noise, α is an over-estimation factor, and β is a flooring factor.

It is essential to estimate the background noise $\tilde{N}(\omega)$ for the spectral subtraction technique. We estimated it based on the distribution of power spectral magnitude for each frequency band [12, 13]. It is assumed that the occurrence of very high amplitude spectra are relatively rare, and they come from the speech signal. The spectral amplitude that is most frequently observed is taken as a mean of noise power at that frequency. The estimated noise is also used to calculate the signal to noise ratio(SNR) to determine the J value of J-RASTA processing [12].

2.3 Harmonic Sieving (HS)

Harmonic structure is one of the most obvious features of the speech signal. However, in most ASR systems, this is smoothed out in order to reduce the variability of the speech spectrum. On the other hand, the harmonic structure has been investigated and used in speech enhancement, speech separation, and computational auditory scene analysis [6, 7, 8, 9]. These techniques can be effective for robust speech recognition, but they are not often evaluated in ASR system with real noise.

We assume that the speech signal in the voiced region exists only in the harmonic structure. With this assumption, we can improve the signal to noise ratio by sieving out the harmonic structure from the mixture of voiced speech and back ground noise. In order to separate out the harmonic structure, we used the method of so called "adaptive comb filtering" as follows:

$$F(t,f) = \begin{cases} A(t,f) & (A(t,f) \text{ is a part of the harmonic structure}) \\ 0 & (\text{otherwise}) \end{cases} \tag{2.4}$$

where $A(t,f)$ is the magnitude spectrum of voiced speech at frequency f on time t.

We can perform this scheme by using pitch detection and voiced-unvoiced detection. Many such algorithms have been investigated in the last few decades; there is a good summary in [14]. We used the subharmonic summation algorithm (SHS) [15] for pitch and voiced/unvoiced detection. In the SHS algorithm, each frame is analyzed by means of the discrete Fourier transform (DFT). The spectra are weighted (higher weight for lower frequency), and resampled in the logarithmic frequency domain with interpolation. Then, the pitch frequency is estimated as the frequency which maximizes the summation of the modified power spectrum on the harmonic structure as follows:

$$f_p = \arg\max_f \left(\sum_{n=1}^{N} W(f)A(nf)\right) \tag{2.5}$$

where N is a integer value which is originally set to 15, $A(f)$ is the magnitude spectrum denoted by DFT, and $W(f)$ is a weight function which reduces the influence of higher frequency spectral components. Instead of taking a maximum value in equation (2.5), we can estimate the pitch frequency from the lattice of the summation by using dynamic time warping [16]. We used the "post editing" algorithm to refine the accuracy of pitch frequency.

The equation (2.4) is rewritten with the output from SHS as follows:

$$F(n,k) = \begin{cases} A(n,k) & (I_w(n,k) \leq C, V(n) \geq 0.5) \\ 0 & (\text{otherwise}) \end{cases} \tag{2.6}$$

where

$$I_w(n,k) = |k - I_p(lf_p(n))|$$

$$I_p(f) = argmin_i(iF_s/N - f)$$

where $A(n,k)$ is the kth frequency bin of N point DFT magnitude spectrum of frame n, $I_w(n,k)$ is the number of frequency bins from k to the closest frequency bin that belongs to a harmonic. $f_p(n)$ is a pitch frequency and $V(n)$ is a voicing confidence at time n; they are calculated by SHS. $V(n)$ takes a value between 0 to 1 and $A(n)$ is likely voiced when $V(n)$ is closer to 1. C is a constant which determines the number of frequency bins for each harmonic frequency. $I_p(f)$ is a frequency index which is closest to the frequency f. The l is an integer value greater than 1, and it is chosen so that $I_p(lf_p(n))$ is close to k.

The quantity $lf_p(n)$ represents the harmonic structure. The l should be limited to less than l_{max} because pitch estimation causes too large an error in harmonic estimation for large l.

Harmonic sieving may cause undesirable spectral distortion in higher frequencies because, as seen in human speech spectrograms, harmonic structure often looses its salient shape at higher frequencies. In order to control the distortion, the maximum frequency of sieving range should be limited not only by l_{max} but also by absolute frequency f_{max}. Thus, the harmonic sieving will performed from pitch frequency f_p to $L(f_p)f_p$. Where $L(f_p)$ is defined as follows:

$$L(f_p) = \min(l_{\max}, f_{\max}/f_p)$$

The equation (2.6) is illustrated in Figure 2.1, which shows DFT spectrograms of word "eight"; a) clean speech, b) speech with +0dB additive noise, and c) harmonically sieved noisy speech (same as (b)). In this example, pitch and voiced/unvoiced variables (i.e. $f_p(n), V(n)$) were derived from clean speech.

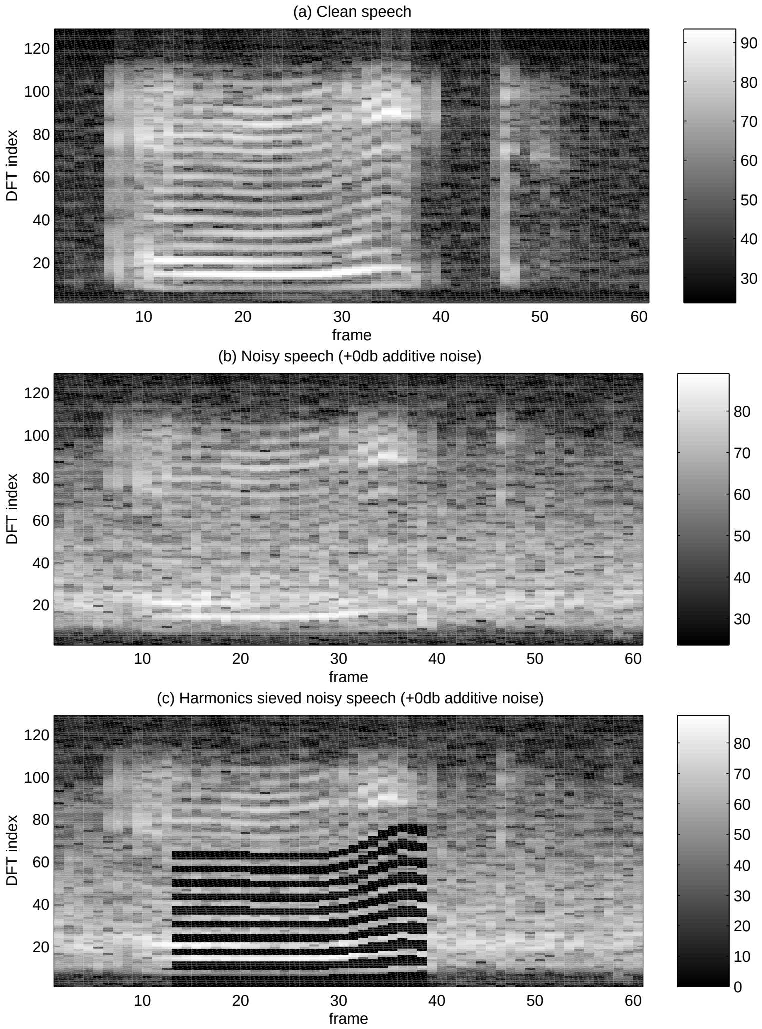

Figure 2.1: Spectrograms of word "eight". (a) Clean speech, (b) Noisy speech (+0dB additive noise), (c) Harmonically sieved noisy speech($l_{max} = 8, C = 1$).

Figure 2.2 shows the spectrogram in the lower frequency region at frame 20 in figure 2.1. It is seen in the noisy spectrum (dotted line) compared with clean spectrum (dashed line) that the frequency regions in between harmonics are filled with noises, while the harmonics

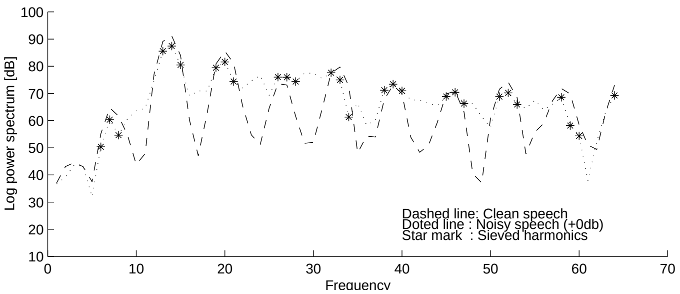

Figure 2.2: Spectrum of clean(solid line), noisy(dotted line), and harmonically sieved noisy speech(asterisk marks).

themselves aren't affected by noise very much. The star marks on the figure show the harmonically sieved noisy speech based on the pitch frequency derived from clean speech. It is seen that, given correct pitch frequency, the sieving process reserve the spectral peaks of the vowel and suppress the noise corrupted region.

In order to compensate the spectrum in higher frequency regions and harmonics, we combine harmonic sieving (equation 2.6) and spectral subtraction (equation 2.3). They are very easily done as follows:

$$F_{ss}(n,k) = \begin{cases} A_{ss}(n,k) & (I_w(n,k) \leq C, V(n) \geq 0.5) \\ 0 & (\text{otherwise}) \end{cases} \tag{2.7}$$

$$A_s s(n,k) = \max(A(n,k) - \alpha \tilde{N}(n,k), \beta A(n,k)) \tag{2.8}$$

where $\tilde{N}(n,k)$ is a estimated noise at kth frequency bin of nth frame. Figure 2.3 shows same spectrogram as shown figure 2.2 but it is preprocessed by the spectral subtraction.

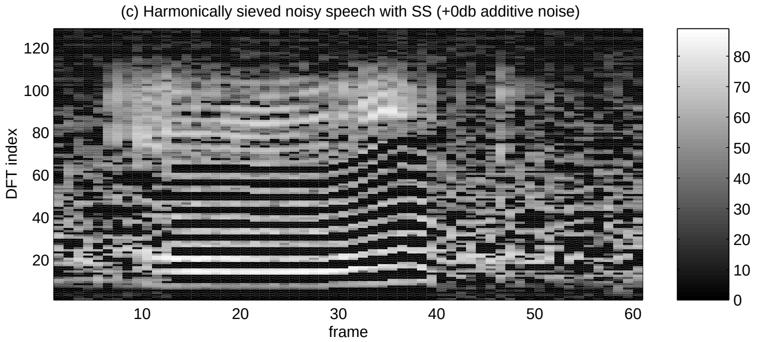

Figure 2.3: Spectrograms of word "eight", Noisy speech (+0dB) additive noise, processed by SS before HS.

Chapter 3

Experiments

3.1 Database

We used an isolated digits database that has been made available by Bellcore. This database has 13 words in the vocabulary: “one”, “two”, “three”, “four”, “five”, “six”, “seven”, “eight”, “nine”, “zero”, “oh”, “yes”, and “no”. They were spoken by 209 speakers through various telephone lines. Since this database is relatively small, we used a “jack-knife method” using 200 speakers (150848 frames), which was divided into four pieces. In each experiment, utterances from 50 speakers were used for testing and the remaining 150 speakers' utterances were used for training. The word error rate reported here is the average of the four experiments.

Noise was added to the clean speech from the digits database. The speech was also filtered to introduce spectral distortion for testing in both additive and convolutional noise environments. The noise source for all experiments except the one reported in section 3.5.2 was car noise, which was recorded[1] over the cellular telephone from a 1978 Volvo 225 running at 55 miles/hour on the freeway with the windows closed.

3.2 The ASR system

The figure 3.1 shows an outline of the ASR system used in our experiments. The feature vector extracted from speech signal is passed to a multi layer perceptron to calculate the a posteriori probability of each phone (56 phones). The phone probabilities are decoded into word probabilities by a Viterbi decoder with lexical and grammar information. Since our task is isolated word recognition, we use a word pair grammar which just assumes a connection from beginning silence to words and from words to ending silence.

3.3 Signal Processing

The speech signal is divided into 25 msec width frames every 10 msec. The spectral subtraction (“SS”) module and harmonic sieving (“HS”) module is attached between the DFT and auditory filter bank of RASTA processing as depicted in figure 3.2. The auditory filter

[1] By Hynek Hermansky

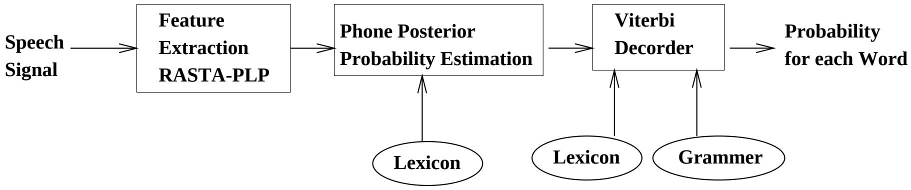

Figure 3.1: The speech recognition system.

is good for smoothing the spectrum which is sieved by HS process. Since the input signal is filtered with a hamming window, and RASTA performs a 256 point FFT for the 25 msec (200 sample) input, the DFT frequency resolution is degraded by more than a factor of two. Therefore, the width of the "teeth" of a comb filter for HS (i.e., C in equation (2.6)) is set to 1 so that 3 frequency bins are sieved for each frequency in the harmonic structure.

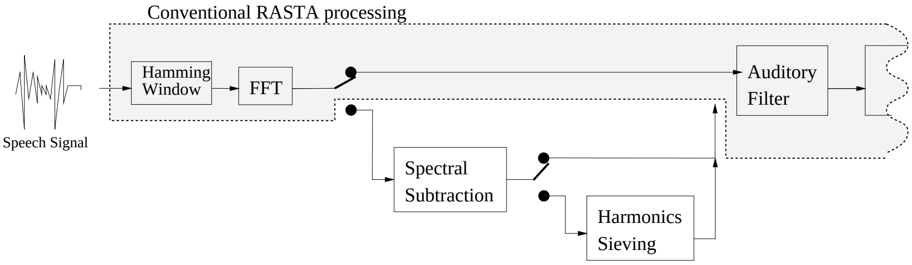

Figure 3.2: J-RASTA processing with spectral subtraction and harmonic sieving module.

3.4 Results

3.4.1 RASTA with Spectral Subtraction

Table 3.1 shows the word error rate for Log-RASTA processing and J-RASTA processing with/without SS. The SS significantly improves the word error rate with J-RASTA processing in the 0dB additive noise environment. In the case of 5dB, J-RASTA is slightly better than J-RASTA with SS, but SS also reduces the word error rate for 5dB additive and convolutional noise. For lower noise, SS does not effect J-RASTA processing. Although SS also improves the result with LogRASTA processing, it still inferior to that of J-RASTA processing. These results show that J-RASTA and SS compensate each other, and SS improves the performance in some very adverse environments. The effective but crude spectral compensation of SS might be smoothed by J-RASTA processing.

Train	Test	clean	+10dB	+10dB	+5dB	+5dB	+0dB
				Filtered		Filtered	
LogRASTA	LogRASTA	2.2	37.3				66.0
	LogRASTA + SS	2.5	31.6	32.0	46.9	47.1	51.8
J-RASTA	JRASTA	2.4	10.0	15.6	18.9	25.4	43.7
	JRASTA + SS	2.3	9.8	14.2	19.6	24.3	30.5

Table 3.1: Word error rate of LogRASTA, J-RASTA, LogRASTA with SS, and J-RASTA with SS ($\alpha = 1.0, \beta = 0.1$).

3.4.2 Effect of flooring factor and overestimation factor

The word error rate for several SS parameters are shown in figure 3.3. Although higher α is effective for 0dB additive noise (solid line), it makes results worse than conventional J-RASTA processing for additive and convolutional noise (dashed line). A typical value for β is 0.1 in the [3, 11], but the word error rate is slightly reduced (less than 7%) by tuning β to 0.3. This might be overtuning to the noise we used.

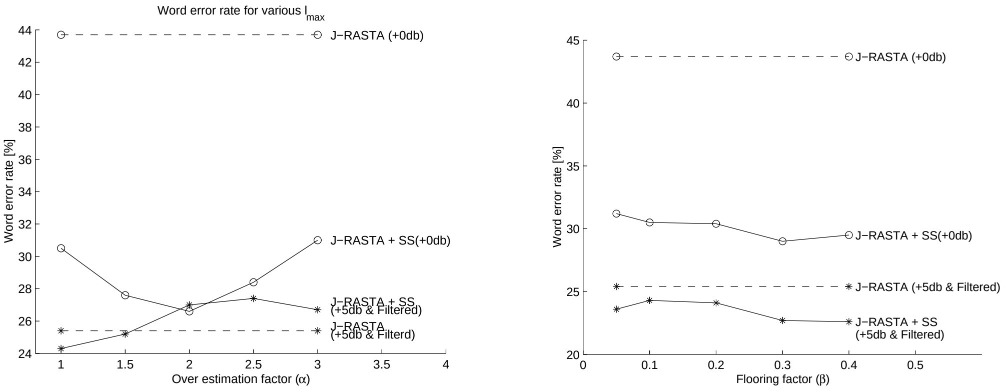

Figure 3.3: Word error rate for various SS parameters (α and β). β was fixed to 0.1 in the left graph, and α was fixed to 1.0 in the right graph.

3.5 J-RASTA with Spectral Subtraction and Harmonic Sieving

In this section, we show the experimental results of J-RASTA processing with SS and HS (i.e., equation 2.7). The parameters for spectral subtraction (α, β) are fixed at $1.0, 0.1$ respectively.

3.5.1 Potential Performance of Harmonic Sieving

First, we derived the pitch frequency from clean speech signal to see the potential of harmonic sieving given correct information[2].

As mentioned in Section 2.3, the number of harmonics to be sieved (i.e. l_{max}) should be limited in order to control the expansion of error of pitch estimation. The word error rate for different l_{max} is plotted in Figure 3.4. The graph shows that l_{max} is necessary. With the relevant l_{max}, The word error rate is reduced by 14% for 0dB additive noise and 20% for 5dB additive and filtered noise when l_{max} is 8.

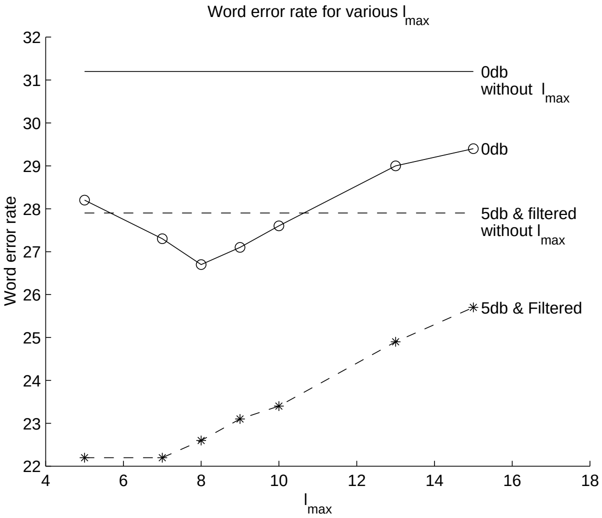

Figure 3.4: Word accuracy according to l_{max} (+0dB additive noise).

In order to see the relationship of the average pitch of each utterance to word error rate, we rearranged results from figure 3.4 according to the average pitch, as shown in Table 3.2. The asterisk mark points to the best result of each column. Higher l_{max} achieves the best result for a lower average pitch, and the best l_{max} is getting lower for a higher average pitch. These results imply the existence of maximum harmonic sieving frequency (i.e., f_{max}) between 600 Hz to 1100 Hz. Thus, we calculated several word error rates according to f_{max}. The results are depicted in figure 3.5. The word error rate is slightly improved with f_{max} around 1000 Hz.

3.5.2 Actual performance

In order to evaluate the actual performance of J-RASTA+SS+HS with real estimated pitch, we ran the SHS over the actual input instead of clean speech then used the output for HS. The results are displayed in Table 3.3. In the experiment, the parameters of SS (α, β) and HS (l_{max}, f_{max}) were set to optimal values from previous sections; $(1.0, 0.1)$,$(8, 1000)$

[2] Although the pitch derived from clean speech also includes errors of around 1% [15], we assume here that it is correct.

	Average Pitch (Low-High) [Hz]							
$l_{\max}$	60-80	80-100	100-120	120-140	140-160	160-180	180-200	200-220
5	28.0	26.7	25.2	26.2	*25.4	*36.0	29.2	*29.9
7	25.8	25.6	23.3	*23.6	*25.4	36.8	*28.8	32.5
8	23.5	24.1	*22.6	24.1	26.3	36.2	30.0	32.5
9	*22.1	*23.8	23.3	24.6	27.8	38.4	30.7	32.5
10	*22.1	24.1	23.9	25.6	28.4	38.4	32.6	35.1
13	*22.1	25.6	26.9	29.2	29.6	40.6	31.8	35.1
15	23.2	26.0	27.5	29.2	30.5	40.1	31.8	35.1
inf	28.6	28.7	28.2	29.7	30.5	41.1	33.0	35.1

Table 3.2: Word error rate according to average pitch for each utterance.

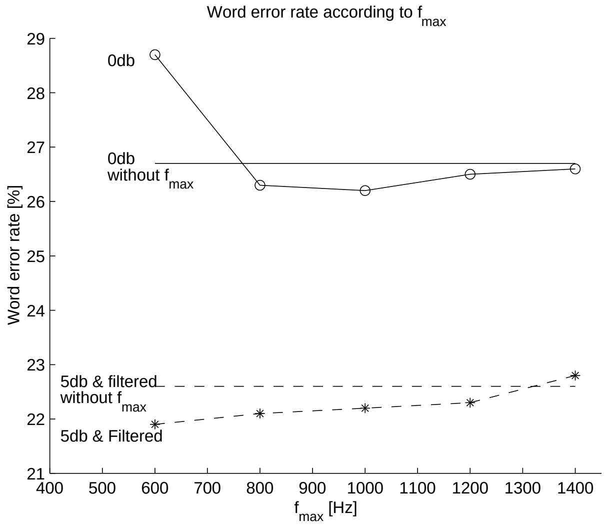

Figure 3.5: Word error rate according to f_{max}.

respectively. The word error rate of J-RASTA and J-RASTA+SS are also displayed again for comparison. The parenthesized number is reduction of error compared with J-RASTA. It can be seen that HS still improves the J-RASTA+SS, even with a realistic pitch estimation error.

	clean	10dB	10dB Filtered	5dB	5dB Filtered	0dB
J-RASTA	2.4	10.0	15.6	18.9	25.4	43.7
J-RASTA+SS	2.3 (4.2)	9.8 (2.0)	14.2 (9.0)	19.6 (-3.7)	24.3 (4.3)	30.5 (30.2)
J-RASTA+SS+HS with correct pitch	2.4 (0.0)	9.6 (4.0)	13.7 (12.2)	18.3 (3.2)	22.2 (12.6)	26.2 (40.0)
J-RASTA+SS+HS with estimated pitch	2.4 (0.0)	9.8 (2.0)	13.8 (11.6)	18.4 (2.6)	23.1 (9.1)	28.5 (34.8)

Table 3.3: Word error rate and percent reduction of error compared to J-RASTA based on pitch estimation with actual input or clean speech ($\alpha = 1.0, \beta = 0.1, l_{max} = 8, f_{max} = 1000$).

We also compared the performances of J-RASTA, J-RASTA+SS, and J-RASTA+SS+HS in several noise environments using the NOISEX-92 database [17]. We used white noise, pink noise, car noise (Volvo), machine gun noise, and Leopard 2 military vehicle noise. Each noise was added to the speech signal with a 0dB SNR. The SNR was calculated based on the energy only over the speech segment. The parameters of SS and HS $\alpha, \beta, l_{max}, f_{max}$ was set to $1.0, 0.1, 8, 1000$ respectively. The car noise is real recorded noise, but its spectrum does not change so frequently, and most of energy is concentrated low frequency region: it is similar to pink noise. The spectrum of noise from Leopard also has its peak at low frequency but the magnitude of the peak is relatively variable over time. Machine gun noise is impulsive noise, so it has a non-stationary wide band spectrum.

	White	Pink	Volvo	Leo	Mac
J-RASTA	73.9	70.3	45.0	71.6	38.4
J-RASTA+SS	44.8	34.4	9.8	29.2	31.1
J-RASTA+SS+HS with correct pitch	41.8	28.6	8.1	32.3	31.3
J-RASTA+SS+HS with estimated pitch	43.9	34.0	8.7	54.7	32.9

Table 3.4: Performance of J-RASTA, J-RASTA+SS, and J-RASTA+SS+HS system in several noise environments."White" is white noise, "Pink" is pink noise, "Volvo" is car noise, "Leo" is Leopard 2 military vehicle noise, and "Mac" is machin gun noise.

Chapter 4

Discussion

The robustness of J-RASTA processing was significantly improved by spectral subtraction, especially in very adverse environments. Furthermore, SS didn't cause bad effects in clean speech or in speech added with lower noise. It might be explained that J-RASTA, which is very flexible but limited in robustness, compensates the SS which is effective but crude. As seen in the experiments with several kind of noises (table 3.4), SS also significantly reduces the error rate for narrow band and quai-stational noise (White, Pink, Volvo, Leo). For wide-band non-stationary noise (i.e., machine gun), it gave a relatively small reduction, because the noise estimator of SS didn't follow the rapid change of the background noise. However, the J-RASTA result for this noise is relatively better than that for other noises. It is understood that RASTA filter suppresses the rapidly changing noise as well.

The error reduction rate for 10db or 5db additive noise with convolutional noise are relatively high. It appears that J-RASTA took an active part in convolutional noise, because additive noises are suppressed by SS and HS.

In our experiments, the flooring parameter of SS was set to 0.1 instead of the 0.3 we had in section 3.4.2 as optimal for noise we used. The optimal parameter improved the results[1] only for the noise used in the tuning, but it makes the results worse for another noises from NOISE-X database. We concluded that the value was overtuned.

Although the error reduction rate of J-RASTA+SS+HS compared to J-RASTA+SS was degraded by a factor of two by using realistic pitch estimation (i.e., SHS) instead of correct pitch, harmonic sieving additively improves the performance. More robust pitch estimation or harmonic structure estimation algorithms are needed to derive the full performance. There are many pitch estimation algorithms, and some may be more robust than SHS. AMPEX(auditory model-based pitch extractor) [18] is another candidate we are considering. We obtained the AMPEX program, but we didn't use it due to time limitations. For harmonic structure estimation, a spectrogram which shows sharp harmonic structure is investigated in [19]. It could be useful for our approach.

Even though the correct pitch is given, the reduction of error by the HS scheme is not extremely large. Two reasons have been considered. First, the signal compensation of HS only covers voiced speech. Non-speech and consonants, which are usually more degraded by noise than voiced speech, are not handled. Since the remaining spectrum

[1] We didn't report them.

after HS should correlate closely with background noise, it can be worthwhile to use it for signal compensation for non-voiced regions. The other reason is that, in this study, we just used the fact that harmonic structure has a comb like shape in the frequency domain. More relevant restricted filtering, for instance the temporal filtering for each tooth of the harmonic structure, may improve the robustness.

Chapter 5

Summary and Conclusion

5.1 Summary

We proposed and evaluated J-RASTA processing with spectral subtraction (J-RASTA+SS) and J-RASTA+SS with harmonic sieving (J-RASTA+SS+HS) using pitch estimation and a voiced/unvoiced decision algorithm. We confirmed that SS improved J-RASTA processing, and that the word error rate was effectively reduced compare to conventional J-RASTA processing in higher noise environments (30% reduction of error in 0db additive noise).

J-RASTA+SS+HS was also evaluated. The parameter to control the HS (l_{max}, f_{max}) was investigated to find an optimal value for our implementation. With these parameters, the harmonic sieving additively improved the performance for some type of noises. (14% of reduction of error given correct pitch, 7% given estimated pitch).

J-RASTA, J-RASTA+SS, and J-RASTA+SS+HS were also evaluated in various noise environments. SS significantly improved the performance. Given correct pitch estimation, HS additively reduced the error rate for white noise (7%) and narrow-band low-frequency noises, i.e. pink noise (16%), car noise (17%). It didn't reduce the error rate for non-stationary noises (Leopard noise, machine gun noise). When the pitch was estimated from noisy input, the reduction in word error rate was degraded. In the case of white noise, pink noise, and car noise the reduction in error rate was 2%, 1.2%, and 11%, respectively. For the last two non-stationary noises, it became worse for Leopard noise (-87%), and machine gun noise (-5%).

5.2 Conclusion

The robustness of J-RASTA processing is significantly improved by spectral subtraction. The noise reduction by harmonic sieving additively improve the performance for narrow band low-frequency noises such as pink noise and car noise. It is expected that good pitch estimation helps not only providing linguistic information but also improving the robustness of ASR system against the background noise.

Chapter 6

Acknowledgements

The SHS pitch estimation algorithm is developed by Dr. Dik Hermes at the Institute for Perception Research, Eindhoven University of Technology(IPO). The author would like to thank Dr. Hermes for providing the SHS program and Dr. Jean-Pierre Martens at ELIS Speech Lab, University of Ghent, Belgium, who kindly provided us with the AMPEX pitch estimation program[1]. The author also would like to specially thank colleagues in the Realization group at ICSI for friendly helping, and Professor Nelson Morgan for the great opportunity for doing research at ICSI.

[1] Because of the limitation of time, we only used SHS, and didn't use AMPEX for experiments.

Bibliography

[1] M. J. F. Gales and S. J. Young. HMM recognition in noise using parallel model combination. In *Proceedings of Eurospeech '93*, pages 837–840, Sept 1993.

[2] M.J.F. Gales and S. J. Young. Cepstral parameter compensation for HMM recognition in noise. *Speech Communication*, 12(3):231–239, 1993.

[3] J.A. Nolazco-Flores and S.J. Young. Continuous speech recognition in noise using spectral subtraction and HMM adaptation. In *ICASSP-94. 1994 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages I–409–I–412. IEEE, 1994.

[4] Hynek Hermansky and Nelson Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing, special issue on Robust Speech Recognition*, 2(4):578–589, Oct. 1994.

[5] Nelson Morgan and Hynek Hermansky. RASTA extensions: Robustness to additive and convolutional noise. In *Proceedings of Workshop on Speech Processing in Adverse Conditions*, pages 115–118, 1992.

[6] R. H. Frazier, S. Samsam, L. D. Bradia, and A. V. Oppenheim. Enhancement of speech by adaptive filtering. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 251–253, April 1976.

[7] Thomas W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *The Journal of the Acoustical Society of America*, 60(4):911–918, - 1976.

[8] J. A. Naylor and S. F. Boll. Techniques for suppression of an interfering talker in co-channel speech. In *ICASSP-87. The 1987 International Conference on Acoustics, Speech, and Signal Processing*, pages 205–208, 1987.

[9] T. Nakatani, M. Goto, and H. Okuno. Localization by harmonic structure and its application to harmonic sound stream segregation. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 653–656, 1996.

[10] Steven F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, ASSP-27(2):113–120, April 1979.

[11] P. Lockwood, J. Boudy, and M. Blanchet. Non-linear spectral subtraction (NSS) and hidden markov models for robust speech recognition in car noise environments. In *ICASSP-92. 1992 International Conference on Acoustics, Speech, and Signal Processing*, pages I–265–I–268. IEEE, Mar. 1992.

[12] Hynek Hermansky and Nelson Morgan. Recognition of speech in additive and convolutional noise based on RASTA spectral processing. In *ICASSP-93. 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, II-83–86, 1993. IEEE.

[13] Hans-Gunter Hirsch. Estimation of noise spectrum and its application to SNR-estimation and speech enhancement. Technical Report TR-93-012, International Computer Science Institute, 1993.

[14] Dik J. Hermes. Pitch analysis. In Martin Cooke, Steve Beet, and Malcolm Crawford, editors, *Visual Representation of Speech Signals*, Wiley professional computing, chapter 1, pages 3–25. Chichester; New York, 1993.

[15] Dik J. Hermes. Measurement of pitch by subharmonic summation. *The Journal of the Acoustical Society of America*, 83(1):257–264, Jan. 1988.

[16] H. Ney. Dynamic programming algorithm for optimal estimation of speech parameter contours. *IEEE Trans. SMC*, 13:208–214, 1983.

[17] A. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones. The NOISEX-92 study on the effect of additive noise on speech recognition. Technical report, DRA Speech Research Unit, Malvern, England, 1992. NoiseX92.

[18] Luc M. Van Immerseel and Jean-Pierre Martens. Pitch and voiced/unvoiced determination with an auditory model. *The Journal of the Acoustical Society of America*, 91(6):3511–3526, June 1992.

[19] Satoshi Imai Toshihiko Abe, Takao Kobayashi. The if spectrogram: A new spectral representation. In *ASVA 97: International symposium on simulation, visualization and auralization for acoustic research and education*, pages 423–429, April 1997.