TecnoLógicas

Noise Reduction in Automatic Detection of Hypernasality in Children

Reducción de Ruido en la Detección Automática de Hipernasalidad en Niños

Helber A. Carvajal-Castaño[1]
Jesús F. Vargas-Bonilla[2]
Claudia V. Isaza-Narvaez[3]

1 Departamento de Ingeniería Electrónica, Universidad de Antioquia, Medellín-Colombia
helber.carvajal@udea.edu.co
2 Departamento de Ingeniería Electrónica, Universidad de Antioquia, Medellín-Colombia
jsvargas@udea.edu.co
3 Departamento de Ingeniería Electrónica, Universidad de Antioquia, Medellín-Colombia
cisaza@udea.edu.co

Tecno. Lógicas., ISSN 0123-7799, Edición Especial, octubre de 2013, pp. 29-42

Abstract

In this paper a methodology to reduce the background noise in a hypernasality detector system using spectral subtraction method is presented, some classical measures of quality and intelligibility are used to evaluate the speech enhancements algorithms used in the system. A linear classifier is used for the hypernasality detection and the results obtained with different spectral subtraction algorithms are compared. The results show that the spectral subtraction techniques can be used to improve the performance of the classifier in the detection of hypernasality when signals are contaminated with additive noise.

Keywords

Noise reduction; quality measures; spectral subtraction; speech enhancement.

Resumen

En este artículo se presenta una metodología para reducir el ruido de fondo en un sistema de detección de hipernasalidad; se utilizan algunas medidas clásicas de calidad e inteligibilidad para evaluar los algoritmos, que mejoran las señales de voz, utilizados en el sistema. La detección de hipernasalidad se realiza con un clasificador lineal y se comparan los resultados obtenidos con diferentes algoritmos de sustracción espectral. Los resultados muestran que las técnicas de sustracción espectral pueden ser usadas para mejorar el rendimiento del clasificador en la detección de hipernasalidad cuando las señales se encuentran contaminadas con ruido aditivo.

Palabras clave

Reducción de ruido; medidas de calidad; sustracción espectral; mejora señales voz.

1. INTRODUCTION

Speech signals are affected by unwanted conditions, such as problems in the transmission channel or additive noise introduced during the reception, which create distortions on the information signal. The main cause of degradation of the speech signals is the presence of background noise, which depends on the characteristics of the environment where the signal is recorded; this affects the quality and intelligibility of the speech signal. Intelligibility refers to a subjective opinion, it depends on the person who is listening, while the quality depends on the percentage of words that can be correctly identified (Vaseghi, 2008).

Distortions caused by background noise generate problems in systems that require speech signal processing, such as speech recognition systems, identification or diagnostic systems. We are interested in hypernasality detection system (Orozco, 2011; Murillo et al.; 2011); this system was developed to work with clean signals captured with professional wiring using frequency sample of 44100Hz and 16 quantization bits (Orozco, 2011). When signals are corrupted by additive noise the system performance decreases, for this reason is necessary to use noise reduction techniques to improve the performance of hypernasality detection system in noisy conditions. In this work the spectral subtraction algorithms are used for noise reduction, this method has been proposed by Boll (1979) and is based on the speech model where the noisy signal can be modeled as the sum of the clean speech signal and additive noise (background noise).

Based in the Boll model, a lot of algorithms has been proposed, in 1979 Berouti et al. (1979) improved the model proposed by Boll to reduce the musical noise, proposing the use of two coefficients to control the spectral subtraction in order to prevent the appearance of peaks in the spectrum (musical noise). Ephraim and Malah (1984) proposed an estimator of the background noise based in the Minimum Mean Square Error (MMSE), a year later used an estimator based in the Log-Spectral Amplitude (Ephrain & Malah, 1985). Based in the Ephraim and Malah estimators, Cappe (1994) performed a demonstration of how the musical noise can be reduced. Wolfe and Godsill (2001) proposed the Maximum a Posteri-

ori (MAP) noise estimator; Gupta et al. (2011) developed a noise robust speech recognition system using spectral subtraction techniques and the Ephraim and Malah estimator.

Different types of spectral subtraction algorithms are compared in this paper: spectral subtraction using a prior signal to noise ratio (SNR) estimation proposed by Scalart et al. (1996) the spectral subtraction using oversubtraction proposed by Berouti et al. (1979) and the multi-band spectral subtraction algorithm proposed by Kamath (2002).

The rest of the paper is organized as follows: in section 2 some spectral subtraction algorithm and measures by evaluate or compare the quality of the enhanced signal are explained, in the section 3 the automatic detection of hypernasality is explained, section 4 contains details about the experiment, section 5 are the results and finally conclusions are presented.

2. SPECTRAL SUBTRACTION METHODS AND QUALITY MEASURES

2.1 Signal Model

When a speaker want to communicate a word, the speaker generates an acoustic signal of the word $s(t)$; this signal may be contaminated by ambient noise and/or distorted by a communication channel or room reverberations, or affected by speaking abnormalities of the talker, and is received as the noisy, distorted and/or incomplete signal $x(t)$ modeled as (1):

$$x(t) = h[s(t)] + n(t) \tag{1}$$

Where the function $h[]$ models the channel distortion, $x(t)$ is the noisy signal, $s(t)$ the clean signal and $n(t)$ is the unwanted additive noise signal. We are assuming that the signal $s(t)$ is only corrupted by additive noise and that the signal $s(t)$ is uncorrelated with the noise signal $n(t)$, (1) changes to:

$$x(t) = s(t) + n(t) \tag{2}$$

In the discrete domain, (2) that is the signal model used in this paper, is:

$$x(k) = s(k) + n(k) \tag{3}$$

This model is valid for hypernasality detection system because the signals have been recorded with professional wiring and the channel distortion is not significant. Then it is considered that only the background noise affect the signal.

2.2 Spectral Subtraction

Spectral Subtraction is one of the most popular (Boll, 1979; Loizou, 2007) methods of reducing the effect of additive noise. This algorithm assumes that $x(k)$, the noisy signal, is composed of the clean speech signal $s(k)$ and the additive noise $n(k)$. Taking the discrete-time Fourier transform on both sides of (3)

$$X(\omega) = S(\omega) + N(\omega) \tag{4}$$

Where *X(ω)*, *S(ω)*, and *N(ω)* are the Fourier transforms of the noisy speech, clean speech and noise signals respectively. In polar form *X(ω)* can be expressed as:

$$X(\omega) = |X(\omega)|e^{j\emptyset_x(\omega)} \tag{5}$$

$|X(\omega)|$ is the magnitude spectrum and $\emptyset_x(\omega)$ is the phase of the noisy speech signal. The noise and the clean signal can be expressed as $N(\omega) = |N(\omega)|e^{j\emptyset_n(\omega)}$ and $S(\omega) = |S(\omega)|e^{j\emptyset_s(\omega)}$ respectively. The magnitude of the noise signal can be replaced by the average magnitude of the noisy signal computed during non-speech activity, and the noisy signal phase can be replaced by the noisy signal phase $\emptyset_x(\omega)$. With this substitution, (4) can be express as (Loizou, 2007):

$$\hat{S}(\omega) = [|X(\omega)| - |\widehat{N}(\omega)|]e^{j\emptyset_x(\omega)} \tag{6}$$

In (6), $\left|\hat{N}(\omega)\right|$ is the magnitude of the noise signal. Finally, the enhanced speech signal can be obtained taking the inverse Fourier transform of $\hat{S}(\omega)$. As the magnitude spectrum of the enhanced signal $\hat{S}(\omega)$ cannot be negative, one solution to this problem is as follows (Loizou, 2007):

$$\hat{S}(\omega) = \begin{cases} |X(\omega)| - \left|\hat{N}(\omega)\right| & if \quad |X(\omega)| > \left|\hat{N}(\omega)\right| \\ 0 & eoc \end{cases} \tag{7}$$

In (6) the expected value of $\left|\hat{N}(\omega)\right|$ is:

$$\left|\hat{N}(\omega)\right| = E\{|N(\omega)|\} \tag{8}$$

With this equation, the spectral error $\epsilon(\omega)$ is defined by (Boll, 1979):

$$\epsilon(\omega) = \hat{S}(\omega) - S(\omega) = |N(\omega)| - \left|\hat{N}(\omega)\right| \tag{9}$$

The goal of the spectral subtraction algorithms is to reduce the spectral error. The more general form of the spectral subtraction algorithm is (Loizou, 2007):

$$\left|\hat{S}(\omega)\right|^p = |X(\omega)|^p - \left|\hat{N}(\omega)\right|^p \tag{10}$$

In (10), when p = 1 is the magnitude spectral subtraction and p = 2 is the power spectral subtraction algorithm.

2.2.1 Spectral subtraction using oversubtraction

In order to remove the peaks that appeared in the spectrum when the spectral subtraction is computed, Berouti et al. (1979) proposed a method that consists of subtracting an overestimate of the noise power spectrum:

$$\left|\hat{S}(\omega)\right|^2 = \begin{cases} |X(\omega)|^2 - \propto \left|\hat{N}(\omega)\right|^2 & if \quad |X(\omega)|^2 > (\propto + \beta)\left|\hat{N}(\omega)\right|^2 \\ \beta\left|\hat{N}(\omega)\right|^2 & eoc \end{cases} \tag{11}$$

When $\propto$ is the oversubtraction factor ($\propto \geq 1$) and β ($0 < \beta \leq 1$) is the spectral floor parameter that is used to reduce musical noise.

2.2.2 Multi-Band spectral subtraction

Kamath and Loizou (2002) proposed a spectral subtraction algorithm based on the fact that the noise will not affect the speech signal uniformly over the entire spectrum, as follows (Loizou, 2007):

$$\left|\hat{S}_i(\omega_j)\right|^2 = \left|\overline{X}_i(\omega_j)\right|^2 - \propto_i \delta_i \left|\widehat{N}_i(\omega_j)\right|^2 \qquad b_i \leq \omega_j < a_i \tag{12}$$

When $\omega_j = 2\pi j/N \quad (j = 0,1,\ldots,N-1)$ are the discrete frequencies, $\left|\widehat{N}_i(\omega_j)\right|^2$ is the estimated noise power spectrum, b_i and a_i are the beginning and ending frequency bins of the ith frequency band, and δ_i is an additional band-subtraction factor that can be individually set for each frequency band.

2.3 Measures of the Quality of Speech

When a spectral subtraction algorithm is used by speech enhancement is necessary the use of the measures by evaluate the quality of the enhanced signal. The distortion measures most used are the following:

2.3.1 Articulation Index (AI)

The AI is defines as (Vaseghi, 2008):

$$AI = \sum_{j=1}^{M} W_j \left(\frac{1}{30}\overline{SNR}(j) + 0.5\right) \tag{13}$$

Where W_j is the band importance function $0 < W_j \leq 1$, M is the number of critical bands and $\overline{SNR}(j)$ are the SNR values $-15\,dB < \overline{SNR}(j) < 15\,dB$. The AI assumes a value between 0 and 1 for SNRs ranging from -15 to 15dB. The AI measure is between 0 to 1, when 1 is when the signal is perfectly intelligible and 0 when the signal is not intelligible.

2.3.2 Itakura-Saito distance (IS)

The IS distance is defined as (Vaseghi, 2008; Loizou, 2007):

$$IS_{12} = \frac{1}{N}\sum_{j=1}^{N}\frac{\left(\boldsymbol{a}_1(j)-\boldsymbol{a}_2(j)\right)^T \boldsymbol{R_1}(j)\left(\boldsymbol{a}_1(j)-\boldsymbol{a}_2(j)\right)}{\boldsymbol{a}_1(j)\boldsymbol{R_1}(j)\boldsymbol{a}_1(j)^T} \tag{14}$$

Where $\boldsymbol{a}_1(j)$ and $\boldsymbol{a}_2(j)$ are the linear coefficient vector from the clean and enhanced signal at frame j and $\boldsymbol{R_1}(j)$ is the autocorrelation matrix obtain from the clean signal. The IS is between 0 to 100, where 0 means that the analyzed signal is equal to the clean signal.

2.3.3 Perceptual Evaluation of Speech Quality (PESQ)

The PESQ is used to predict subjective opinion scores of a degraded speech sample (Vaseghi, 2008; Loizou, 2007) and is computed as a linear combination of the average disturbance value D_{ind} and the average asymmetrical disturbance values A_{ind} as (Hu & Loizou, 2008):

$$PESQ = 4.5 - 0.1D_{ind} - 0.0309A_{ind} \tag{15}$$

The PESQ score is 0.5 to 4.5, where 4.5 means that the analyzed signal is very good.

2.3.4 Loglikelihood Ratio (LLR)

$$LLR_{12} = log_{10}\left(\frac{\boldsymbol{a}_2(j)\boldsymbol{R_1}(j)\boldsymbol{a}_2(j)^T}{\boldsymbol{a}_1(j)\boldsymbol{R_1}(j)\boldsymbol{a}_1(j)^T}\right) \tag{16}$$

Where $\boldsymbol{a}_1(j)$, $\boldsymbol{a}_2(j)$ and $\boldsymbol{R_1}(j)$ are defined as in the IS distance. The LLR is between 0 to 100, similar to the IS.

3. HYPERNASALITY DETECTION SYSTEM

Based on the acoustical characteristics of the voice, a system to diagnose hypernasality was developed (Murillo et al., 2011; Orozco et al., 2011). However, this system works only with signals record-

ed under ideal conditions, namely in an isolated booth. The hypernasality detection was based in acoustic, noise, cepstral and nonlinear dynamic features. The basic steps are the follows:

3.1 Acoustic Analysis

In this part some acoustic characteristics are taken, as: Jitter, Shimmer, Harmonic to Noise Ratio (HNR), Normalized Noise Energy (NNE), Harmonic to Noise Ratio in Cepstral domain (CHNR), Glottal to Noise Excitation Ratio (GNE) and eleven Mel Frequency Cepstral Coefficient (MFCC). Whit this features the classification is performed.

3.2 Features Selection and Classification

In (Murillo et al., 2011) the features selection is made using Principal Component Analysis (PCA). Finally, with the principal components, the classification was performed using the Linear-Bayes classifier.

4. EXPERIMENTAL SETUP

4.1 Database

The data base used for tested this methodology was provided by Grupo de Procesamiento y Reconocimiento de Señales (GPRS) from the Universidad Nacional de Colombia, Manizales. This data base contains recording of five Spanish vowels pronounced by children aged between 5 and 15. The database contains 266 registers, 156 of them were labeled as hypernasal by a phoniatry expert and the rest 110 were labeled as healthy (Orozco, 2011).

4.2 Experiment

The database was contaminated with additive noise with a SNR of 3, 5, 10 and 20dB, on this database are calculated the acoustic characteristics of each signals as explained in (Orozco,

2011). The SNR of a signal recorded in a phoniatry office is approximately 10dB. The SNR = 3dB simulates a highly contaminated signal, while a SNR = 20dB simulates a clean signal recorder in a controlled environment. The enhanced database is obtained by applying spectral subtraction algorithms to contaminated database.

Since in the real conditions the noise present in a phoniatry office is approximately 10dB, to train the classifier, we used enhanced signal which are obtained by applying the spectral subtraction algorithm over the signals that were contaminated with a SNR = 10dB. The classifier is tested with the clean signals, the contaminated signals with a SNR of 3, 5, 10 and 20dB and the enhanced signals with a SNR of 3, 5 and 20dB.

5. RESULTS

Table 1 shows the results of the test over contaminate and enhanced signals, when the Scalart spectral subtraction algorithm is used. Training performance when the classifier is trained with the signals obtained by applying the Scalart and Filho (1996) algorithm is 89.62 ± 0.42, the results show an improvement in the hypernasality detection when the SNR of the signals is 3, 5 and 10 dB. When the SNR = 20dB, the results with the corrupted and enhances signals is comparable; the standard deviation of the enhanced signal contains the corrupted signal.

Table 1. Classification performance [%] when Scalart Algorithm is used

	SNR			
	3 dB	5 dB	10 dB	20 dB
Corrupted	65,65 ± 9,97	76,32 ± 9,62	89,06 ± 2,08	87,83 ± 1,66
Enhanced	72,73 ± 0,83	88,14 ± 0,95	**89,62 ± 0,42**	86,09 ± 2,14

The Table 2 shows the test results over the corrupted and enhanced signals when the Berouti et al. (1979) algorithm is used. The classifier performance when this algorithm is used is 89.06 ± 3.34. The results show, as in the previous, an improvement in the hypernasality detection when the signals are enhanced.

Table 2. Classification performance [%] when Berouti Algorithm is used

	SNR			
	3 dB	5 dB	10 dB	20 dB
Corrupted	54,73 ± 8,50	65,09 ± 10,3	88,09 ± 2,21	89,88 ± 0,79
Enhanced	73,06 ± 3,97	89,72 ± 1,33	**89,06 ± 3,34**	89,34 ± 1,14

The classifier performance when the multi-band (Kamath & Loizou, 2002) algorithm is used is 83.81 ± 1.01. The Table 3 show the results when this algorithm is used.

Table 3. Classification performance [%] when Kamath Algorithm is used.

	SNR			
	3 dB	5 dB	10 dB	20 dB
Corrupted	62,48 ± 4,15	68,23 ± 6,12	84,76 ± 3,87	86,64 ± 2,90
Enhanced	70,36 ± 1,17	75,47 ± 2,01	**83,81 ± 1,01**	86,49 ± 2,37

The results obtained when the classifier is tested with clean signals are shown in the Table 4. These results are similar to those obtained when the classifier is tested with signals with a SNR = 10 dB; therefore, it is possible to obtain a good hypernasality detection using spectral subtraction techniques when signals are recorded with background noise.

Table 4. Classification using the Clean Signals

Algorithm used for Spectral Subtraction	Performance
Scalart	75,37 ± 2,80
Berouti	**84,82 ± 1,65**
Kamath	74,82 ± 4,27

Table 5 shows the measures of the quality of the speech calculated on the noisy and enhanced signals when the selected spectral subtraction algorithms are used, this table shows, in general, improvement in the quality measures used.

The results show that the Berouti algorithm have the best performance with respect to other spectral subtraction techniques evaluated, as seen in Table 2, where percentages are higher compared with the results obtained with other algorithms, also, the classification test with the clean signals is better than that obtained with Scalart and Kamath algorithm.

Table 5. Quality Measures

SNR	Measure	Corrupted	Enhanced Scalart	Enhanced Berouti	Enhanced Kamath
3 dB	IS	3,39 ± 0,83	1,48 ± 0,82	1,14 ± 0,61	1,91 ± 0,96
	LLR	1,44 ± 0,63	0,80 ± 0,51	0,87 ± 0,50	1,02 ± 0,56
	PESQ	2,49 ± 0,32	2,62 ± 0,35	1,99 ± 0,32	1,71 ± 0,45
	AI	0,07 ± 0,02	0,73 ± 0,06	0,70 ± 0,04	0,71 ± 0,01
5 dB	IS	3,05 ± 0,81	1,15 ± 0,61	0,91 ± 0,51	1,45 ± 0,85
	LLR	1,24 ± 0,60	0,71 ± 0,47	0,66 ± 0,41	0,78 ± 0,47
	PESQ	2,68 ± 0,32	2,83 ± 0,45	2,44 ± 0,63	3,04 ± 0,41
	AI	0,12 ± 0,02	0,79 ± 0,05	0,74 ± 0,07	0,76 ± 0,04
10 dB	IS	2,30 ± 0,71	0,75 ± 0,46	0,80 ± 0,42	1,00 ± 0,55
	LLR	0,80 ± 0,50	0,37 ± 0,31	0,34 ± 0,30	0,52 ± 0,25
	PESQ	3,19 ± 0,35	3,35 ± 0,52	2,78 ± 0,77	3,49 ± 0,27
	AI	0,24 ± 0,04	0,92 ± 0,04	0,88 ± 0,03	0,86 ± 0,06
20 dB	IS	1,23 ± 0,44	0,48 ± 0,20	0,82 ± 0,31	1,01 ± 0,41
	LLR	0,22 ± 0,27	0,10 ± 0,10	0,10 ± 0,18	0,36 ± 0,18
	PESQ	4,00 ± 0,24	3,85 ± 0,40	2,92 ± 0,84	3,67 ± 0,25
	AI	0,48 ± 0,07	0,98 ± 0,01	0,96 ± 0,01	0,95 ± 0,01

The method chosen for spectral subtraction is the proposed by Berouti et al. (1979), according to the results, presents the best performance and delivering results comparable to those obtained with clean signals.

6. CONCLUSIONS

In this paper was proposed the use of some spectral subtraction algorithms to improve the performance of a hypernasality detection system when the signals are contaminated by additive noise. The results show that using the Berouti et al. algorithm, to improve speech signals, a performance comparable with the result obtained when using clean signals is obtained.

The measures of quality and intelligibility indicate how well the spectral subtraction algorithms work. The results show that the spectral subtraction algorithms used enhanced the signals compared with the corrupted signals, allowing improved perfor-

mance of the hypernasality detection system when the signals are corrupted by additive noise.

When the signal to noise ratio of the signals is greater that 10dB is not appropriate the use of spectral subtraction algorithms because the spectral subtraction degrades the signal instead of enhanced the signals. Proposed methodology allows the classification of hypernasality when speech signals are contaminated obtaining similar results that using clean speech signals. Thus, is no longer need to record the signals with expensive equipment to get a good classification. In the future, is expected that the system of hypernasality detection can be implemented at low cost.

7. ACKNOWLEDGEMENTS

This work was supported by the CODI, Universidad de Antioquia (COL).

8. REFERENCES

Berouti, M., Schwartz, R. & Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. Acoustic, Speech and Signal Processing. ICASSP79 (Vol. 4, pp 208-211).

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. Acostic, Speech and Signal Processing, IEEE Transactions on, vol. 27, no. 2, pp. 113-120.

Cappe, O. (1994). Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. Acoustic, Speech and Signal Processing. IEEE Transactions (Vol. 2, no 2, pp. 345-349).

Cohen, I. (2004). Speech enhancement using a noncausal a priori SNR estimator. Signal Processing Letters, IEEE (Vol. 11, no. 9, pp. 725-728).

Ephraim, Y. & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. Acoustic, Speech and Signal Processing. IEEE Transactions (Vol. 32, no. 6, pp. 1109-1121).

Ephraim, Y. & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. Acoustic, Speech

and Signal Processing. IEEE Transactions (Vol. 33, no. 2, pp. 443-445).

Gupta, V. K., Bhowmick, A., Chandra, M. & Sharan, S. N. (2011). Speech enhancement using MMSE estimation and spectral subtraction methods. Devices and Communication (ICDeCom), 2011 International Conference, pp 15.

Kamath, S. & Loizou, P. C. (2002). A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. Proc. IEEE Int. Conf. Acoustic, Speech and Signal Processing.

Loizou, P. C. (2007). Speech Enhancement, Theory and Practice, 1ed.

Loizou, P. C. & Hu, Y. (2008). Evaluation of objective quality measures for speech enhancement. Audio, Speech and Language Processing, IEEE Transactions (Vol. 16, no. 1, pp. 229-238).

Murillo, S., Orozco, J. R., Vargas, J. F., Arias, J. D. & Castellanos, C. G. (2011) Automatic detection of hypernasality in childrens. Proceeding of the 4th international conference on Interplay between natural and artificial computation: new challenges on bioinspired applications, Volume Part II, pp. 167-174.

Orozco, J. R. (2011) Análisis acústico y de dinámica no lineal para la detección de hipernasalidad en la voz. MSc Thesis, Universidad de Antioquia.

Orozco J. R., Murilo, S., Alvares, A. M., Arias, J. D., Delgado, E., Vargas, J. F., & Castellanos, C. G. (2011). Automatic selection of acoustic and non-linear dynamic features in voice signals for hypernasality detection. INTERSPEECH 2011.

Scalart, P. & Filho, J. (1996). Speech enhancement based on a priori signal to noise estimation. Proc. IEEE Int. Conf. Acoustic, Speech and Signal Processing, 629-632.

Vaseghi, S. V. (2008). Advanced Digital Signal Processing and Noise Reduction, 4 ed.

Wolfe, P. J. & Godsill, S. J. (2001). Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement. Statistical Signal Processing. Proceeding of the 11th IEEE Signal Processing Workshop, 496-499.