Source Separation via Spectral Masking for Speech Recognition Systems

Gustavo Fernandes Rodrigues, Thiago de Souza Siqueira, Ana Cláudia Silva de Souza, and Hani Camille Yehia

Abstract*—In this paper we present an insight into the use of spectral masking techniques in time-frequency domain, as a preprocessing step for the speech signal recognition. Speech recognition systems have their performance negatively affected in noisy environments or in the presence of other speech signals. The limits of these masking techniques for different levels of the signal-to-noise ratio are discussed. We show the robustness of the spectral masking techniques against four types of noise: white, pink, brown and human speech noise (bubble noise). The main contribution of this work is to analyze the performance limits of recognition systems using spectral masking. We obtain an increase of 18 % in the speech hit rate, when the speech signals were corrupted by other speech signals or bubble noise, with different signal-to-noise ratio of approximately 1, 10 and 20 dB. On the other hand, applying the ideal binary masks to mixtures corrupted by white, pink and brown noise, results in an average growth of 9 % in the speech hit rate, with the same differents signal-to-noise ratios. The experimental results suggest that the spectral masking techniques are more appropriate when applied to bubble noise, which is produced by human speech, than to white, pink and brown noise.

***Keywords*—Blind source separation, Independent component analysis, Neural networks, Spectral masking, Speech recognition.**

I. Introduction

When several people are talking at the same time in a meeting or public places, it is necessary to separate the voice of a given person or an specific source from other interference sources so that each speaker can be recognized. Independent components analysis has been an important source separation technique, however, with the presence of noise and reverberation, the separated signals have strong residual components of other interference sources [1]. In these cases, a signal preprocessing method must be used in order to reduce other sources of interference. Our goal is to show the potential improvement of automatic speech recognition in noisy environments or with multiple speech signals. In this paper, we show that spectral masking techniques, used as preprocessing tools, reduce other sources of interference and increase the efficiency of the speech recognition systems.

Several works use observation vectors of uncertainties in the decoding process for the treatment of noisy signals in the automatic speech recognition task [2]–[4]. When dealing with speech recognition in environments with several speakers, such as the ones in [5]–[7], some authors suggest the use of binary masking [8], [9]. However, when speech signals are exposed to environments with reverberation and in the presence of other speakers, the process of extracting the masks becomes extremely difficult. This paper intends to quantify and analyze the efficiency of the spectral masking in speech recognition tasks.

The majority of speech recognition systems does not show good performance in noisy environments or when there are interferences from other voices. Therefore, we intend to improve the efficiency of speech recognition systems through the implementation of a source separation method using spectral masking in time-frequency domain, as a preprocessing stage. Time-frequency domain masking is used to extract an specific speech signal from the noise corrupted signal [10], [11]. The mel-cepstral parameters are used in the speech recognition step to provide the input data to the speech recognition system. Some papers show that binary masking provides extracting information in time-frequency domain which best characterizes speech signal [11]. The necessity to improve the speech recognition system's performance in environments under adverse conditions and multiple speakers has attracted researchers attention and many papers about separation of speech signals have been published [1], [4]–[6], [9], [10]–[14].

This paper is organized as follows. In Section II we discuss source separation techniques via spectral masking. In Section III we describe the steps to obtain the speech signals parameters and discuss the implementation of the speech recognition system. In Section IV we show the results of the experiments and simulations done to verify the influence of noise and other voices in speech recognition tasks. The tests made to analyze the limits and improvement capacity of the speech recognition systems through spectral masking as well as the analysis of the results obtained are detailed in Section IV. Finally, Section V outline the conclusions of this work.

II. Spectral Masking in time-frequency domain

An specific sound source can be recovered by applying a weighted mask to an acoustic mixture at each point in the time-frequency domain. The regions dominated by this source receive higher weights than the ones where other sound sources of the analyzed mixture prevails. The masks may be binary or assume real values. The use of a binary mask is motivated by the masking process which occurs in human audition, where a more intense sound can mask or obscure a less intense one within the same critical band. With the

G. F. Rodrigues, T. S. Siqueira and A. C. Souza are Department of Telecommunications and Mechatronic Engineering, Federal University of São João del-Rey, Ouro Branco, Minas Gerais, Brazil (corresponding author to provide phone and fax: +55 31 37413583, e-mail: gustavofernandes@gmail.com).

H. C. Yehia is with Department of Electronic Engineering, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.

Manuscript received October 28, 2012.

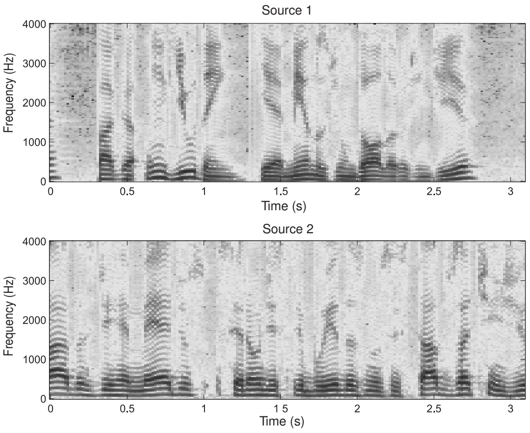

Fig. 1. Two speech spectrograms.

purpose of separating the voice signals, [15] proposed the usage of an ideal binary mask. Given a speech signal $s(t,f)$ and a noise $n(t,f)$, where t and f represent the instant in time and frequency, respectively, the ideal binary masking $m(t,f)$ can be obtained through the following expression:

$$m(t,f) = \begin{cases} 1 & if \quad s(t,f) > n(t,f), \\ 0 & \textit{otherwise.} \end{cases} \tag{1}$$

A similar approach was adopted in [16], who observed an orthogonal tendency in different voice signals in time-frequency domain with high resolution, showing that it is possible to separate signals through binary masking. Several papers have shown that the speech signal reconstructed by an ideal binary mask is intelligible when extracted from a mixture of two or more speakers [9], [17], [18].

The ideal binary mask considered here is a binary matrix which assumes value one when the signal energy is stronger than the interference signal for a specific frequency in a given instant and assumes value zero otherwise. When we apply an ideal mask to an instantaneous mixture of two speech signals, we notice that the signal obtained is perfectly audible, with good quality and no interference from the other signal.

In order to obtain the ideal mask, the source signals (s_1 and s_2) are transformed to time-frequency domain and their spectrograms are given, respectively, by:

$$\begin{aligned} s_1(t) \rightarrow S_1(w,t), \\ s_2(t) \rightarrow S_2(w,t), \end{aligned} \tag{2}$$

where w represents angular frequency and t represents the time instant of the voice frame being analyzed. The binary masking can be determined by comparing the magnitude of two spectrograms, as shown in Fig. 1.

The ideal masks (M_1 and M_2) were obtained as follows:

$$\begin{aligned} M_1(w,t) = 1, \quad for \quad |S_1(w,t)| > |S_2(w,t)|, \\ M_2(w,t) = 1, \quad for \quad |S_2(w,t)| > |S_1(w,t)|, \end{aligned} \tag{3}$$

and the other values of the masks are equal to zero.

Fig. 2 shows an example of ideal binary masks applied to two speech signals, sampled at 8000 Hz, divided in frames of 512 samples with 50% overlap.

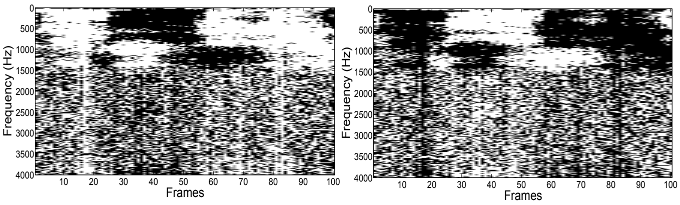

Fig. 2. Images of binary ideal masks from the sources (s_1 and s_2) for 100 speech frames.

Spectral masking techniques can be applied to speech signal separation, specially when the mixture is corrupted by noise signals. When applying an ideal binary mask to separate two speech signals, it is possible to improve the signal-interference ratio over 40 dB (SIR), but it also increases the signal-distortion ratio (SDR) in 20 dB [1]. However, an ideal binary mask can not be obtained without knowledge of the real signals and an approximation is needed. In order to obtain this approximation, several methods to estimate masks based on ICA (independent component analysis) can be found in the literature, both for binary and continuous masks [12], [14].

The signal-distortion ratio decreases about 3 dB when approximately 10% of the bits of the ideal binary mask are inverted, as shown in [19]. An error of 10% of the bits in the binary masks estimate is acceptable without loss of intelligibility [19], [20].

The objective of this paper is to show the potential improvement of the speech recognition systems, in noisy environments or with multiple speech signals, using spectral masking techniques. We present a signal preprocessing method to reduce other sources of interference, using ideal binary masks. In this experiment, the speech signals or noise signals were known. This fact allows to obtain the ideal binary mask from each mixture analyzed. We use all the recordings to simulate mixtures corrupted by noise, and therefore to obtain the ideal masks. The ideal binary masks obtained were applied to a mixture corrupted by other speech signal or noise, sampled at 8000 Hz, divided in frames of 512 samples with 50% overlap. The ideal binary masks were used to separate the signal of interest (speech signal) from the noise, as a preprocessing step for the speech signal recognition.

The ideal masks obtained were directly applied to the speech signal mixtures in time-frequency domain, before the extraction of the mel-cepstral coefficients and the learning process of the neural network used in the speech recognition tests. The ideal masks were not applied in the training set used to train the neural network.

III. Speech Recognition System

Speech recognition systems have low performance in noisy environments or in the industry. In this work, we are concerned with real situations where speech signals are corrupted by noise, including other speech signals. A speaker dependent speech recognition system was used to recognize isolated voice commands from a limited vocabulary (30 Portuguese words) and each word was recorded 20 times by one speaker. The corpus consists of voice commands that can be used in automation and control systems . The Portuguese words used were: right, left, stop, go back, go on, ahead, behind, on, off, fast, slow, turn on, turn off, up, down, speed up, lock, unlock and alarm. Besides these words, the database consists of eleven digits: 0 to 10. The ideal binary masks were used to separate the target speech signals from a mixture with other speech signal or other types of noise. We use all the recordings to simulate mixtures corrupted by noise, and therefore to obtain the ideal masks. The ideal masks obtained were directly applied to the speech signal mixtures in time-frequency domain, before the extraction of the mel-cepstral coefficients and the learning process of the neural network used in the speech recognition system.

A. Extraction of Mel-Cepstral coefficients

The mel-cepstral parameters are the most used features in speech recognition systems as input data. A Mel is unity of measurement for the perceived pitch of a tone. It does not have a linear correspondence to the physic frequency. The Mel scale is defined by a mapping between the frequency scale (in Hz) and the perceived frequency scale (in Mel). The mapping is linear until approximately 1 KHz and logarithmic for superior frequencies. The frequency scale denominated Mel is closely related to the critical-band of the auditory human system and the mel-cepstral coefficients are obtained from the Mel frequency.

The extraction of the mel-cepstral coefficients consists of the following the steps: (i) to obtain the magnitude spectrum of the signal by applying Fourier transform; (ii) to calculate each output of a filter bank in the Mel scale as a sum of the weighted spectral magnitude of each frame; (iii) to obtain the logarithm of the magnitude spectrum in each filter output; (iv) to take the discrete cosine transform (DCT) from each frame.

The speech signals used were recorded with a sampling frequency of 8000 Hz. In the extraction step, the data were divided in frames of 512 samples with 50% overlap. For each frame, a Hamming window was applied followed by 13th order mel-cepstral coefficients.

B. Principal Components Analysis

Principal Components Analysis (PCA) consists of a linear transformation of "*m*" original variables in "*m*" new variables, in such a way that the first new variable accounts for as much of the variability in the data as possible and each succeeding component in turn has the highest variance possible under the constraint that it is uncorrelated with the preceding ones, until each variance in the set has been explained. The purpose of this technique in this case is to allow a reduction in the dimension of the data, therefore minimizing error.

For each frame we have used 13 mel-cepstral coefficients and therefore the data set was represented by a Nx13 vector, where N corresponds to the number of frames in analysis.

$$x_n = [x_1(n)\ x_2(n)\ \cdots\ x_{13}(n)]^T, \tag{4}$$

where $[.]^T$ denotes a transpose matrix. The variable x_n, $1 \leq n \leq N$ represents one frame of the audio signal being analyzed. The set of all frames is represented by the N-sized vectors:

$$X = [x_1\ x_2\ \cdots\ x_N], \tag{5}$$

where each column of X denotes the 13 coefficients of each signal frame. From that, the covariance matrix is defined by:

$$C = [X - \mu][X - \mu]^T, \tag{6}$$

where μ is a mean vector. Then, through the decomposition of singular values (SVD) one can denote the covariance matrix as:

$$C = USU^T, \tag{7}$$

where U is a matrix whose columns are eigenvectors of C and S is a diagonal matrix containing the respective eigenvalues of C.

The sum of the eigenvalues represents the total variance observed in C. Therefore, if the sum of the first k eigenvalues reaches a proportion, as the one considered, of 85% of the sum of all eigenvalues, then the first k eigenvectors of C will account for most of the total variance observed in the data set. In this paper we used the first three components as the input vector for the pattern classification by neural network. The dimensions of speech features were reduced to a vector of 39 values for each word.

C. Learning with neural network

A training corpus of 600 utterances was used from which half was used for training and the remaining for testing. The input speech was reduced to a vector of 39 values for each word from the mel-cepstral coefficients (the dimension was reduced by PCA).

Algorithms based on neural network of type multi-layer perceptron, using backpropagation algorithm to the supervised learning process has been used in voice recognition systems [21]–[24]. The network used in this paper is a feedforward multilayer perceptron (MLP) trained with stochastic back-propagation algorithm. This experiment uses three layer MLP: one input layer, one hidden layer and one output layer. The feature vectors representing speech pattern are fed into neural network at the input layer. Only 39 values (from mel-cepstral coefficients with the dimension reduced by pca) for each word are fed to the neural network. In this neural network there is a single hidden layer that has 30 neurons. The numbers of neurons in output layers is set to five. The output of the network is a binary value representing the recognized word. The hidden and output neurons are activated using sigmoidal and linear activation functions respectively. Once the network is created, it can be trained for a specific problem by presenting training inputs and their corresponding targets (supervised training). A set of 10 samples of each word (300 utterances) was used as training data and another part as test data. The binary masks were not used in the training set. The ideal masks were applied only in the test set used to test the neural network.

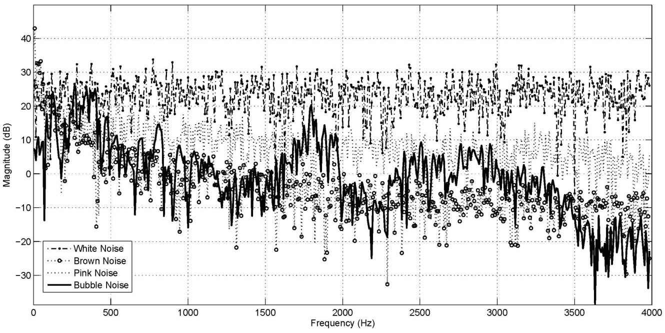

Fig. 3. Frequency spectrum of white, pink, brown and bubble noise.

TABLE I
THE SPEECH SIGNALS WERE CORRUPTED BY ONE SPEECH SIGNAL FROM ANOTHER SPEAKER WITH DIFFERENT LEVELS OF SIGNAL-TO-NOISE RATIO (1, 10, 20 E 30 DB). THE TESTS WERE PERFORMED USING AN IDEAL BINARY MASK AND WITHOUT APPLICATION OF THE IDEAL MASK.

Mixture corrupted by one speech signal					
Without masking			With masking		
SNR (dB)	Hit Rate (%)	Standard Deviation	SNR (dB)	Hit Rate (%)	Standard Deviation
1	31	9	1	48	11
10	43	12	10	64	15
20	66	16	20	79	16
30	85	19	30	87	17

IV. RESULTS

The recordings were made in a laboratory with low noise level, with approximately zero reverberation time and sampling frequency of 8000 Hz. The hit rate (word recognition rate) was used as a measure of the speech recognition efficiency. Another common metric is the word error rate (WER). We decided to use the hit rate measure as the mirror representation of the word error rate. In order to obtain the speech hit rates, 100 simulations were made for each case being analyzed. Different simulations were done for testing, where the original speech signal was corrupted by other speech signals from different speakers and by different types of noise. In each case different levels of signal-to-noise ratio (1, 10, 20 and 30 dB) were considered.

The definition of noise is derived from a random signal, but it can have different characteristic statistical properties. In this paper, we analyzed the limits of spectral masking techniques for the following types of noise: white, pink, brown and a human speech noise (bubble noise). By definition, the white noise has a flat frequency spectrum. The Pink noise or "$1/f$ noise" is a signal with power spectral density inversely proportional to the frequency. The power density, compared with white noise, decreases by 3 dB per octave (density is proportional to $1/f$). The brown noise refers to a power density which decreases 6 dB per octave with increasing frequency (density is proportional to $1/f^2$). In order to create a noise with a frequency spectrum similar to the human speech we concatenated all the words of the vocabulary used is this work, spoken by 3 males and 3 females [25]. The frequency spectrum of these different types of noise used is shown in 3.

TABLE II
THE SPEECH SIGNALS WERE CORRUPTED BY TWO SPEECH SIGNALS FROM OTHER SPEAKERS WITH DIFFERENT LEVELS OF SIGNAL-TO-NOISE RATIO (1, 10, 20 E 30 DB). THE TESTS WERE PERFORMED USING AN IDEAL BINARY MASK AND WITHOUT APPLICATION OF THE IDEAL MASK.

Mixture corrupted by two speech signal					
Without masking			With masking		
SNR (dB)	Hit Rate (%)	Standard Deviation	SNR (dB)	Hit Rate (%)	Standard Deviation
1	28	9	1	48	12
10	47	12	10	65	18
20	60	18	20	81	18
30	85	15	30	88	19

TABLE III
THE SPEECH SIGNALS WERE CORRUPTED BY THREE SPEECH SIGNALS FROM OTHER SPEAKERS WITH DIFFERENT LEVELS OF SIGNAL-TO-NOISE RATIO (1, 10, 20 E 30 DB). THE TESTS WERE PERFORMED USING AN IDEAL BINARY MASK AND WITHOUT APPLICATION OF THE IDEAL MASK.

Mixture corrupted by three speech signal					
Without masking			With masking		
SNR (dB)	Hit Rate (%)	Standard Deviation	SNR (dB)	Hit Rate (%)	Standard Deviation
1	32	11	1	56	12
10	52	16	10	68	16
20	68	21	20	82	16
30	89	15	30	89	17

TABLE IV
THE SPEECH SIGNALS WERE CORRUPTED BY DIFFERENT TYPES OF NOISE WITH DIFFERENT LEVELS OF SIGNAL-TO-NOISE RATIO (1, 10, 20 E 30 DB). THE TESTS WERE PERFORMED USING AN IDEAL BINARY MASK AND WITHOUT APPLICATION OF THE IDEAL MASK.

Mixture corrupted by a white noise					
Without masking			With masking		
SNR (dB)	Hit Rate (%)	Standard Deviation	SNR (dB)	Hit Rate (%)	Standard Deviation
1	23	9	1	33	10
10	30	9	10	51	11
20	53	9	20	52	12
30	64	16	30	72	17

Mixture corrupted by a pink noise					
Without masking			With masking		
SNR (dB)	Hit Rate (%)	Standard Deviation	SNR (dB)	Hit Rate (%)	Standard Deviation
1	25	9	1	37	9
10	44	12	10	54	13
20	58	17	20	63	16
30	77	21	30	79	17

Mixture corrupted by a brown noise					
Without masking			With masking		
SNR (dB)	Hit Rate (%)	Standard Deviation	SNR (dB)	Hit Rate (%)	Standard Deviation
1	54	15	1	65	16
10	56	16	10	66	15
20	78	20	20	80	16
30	88	17	30	91	18

Mixture corrupted by a bubble noise					
Without masking			With masking		
SNR (dB)	Hit Rate (%)	Standard Deviation	SNR (dB)	Hit Rate (%)	Standard Deviation
1	32	8	1	53	14
10	60	16	10	74	13
20	77	21	20	81	14
30	89	14	30	90	13

The following cases were analyzed: i) speech signal corrupted by a speech signal from another speaker; ii) speech

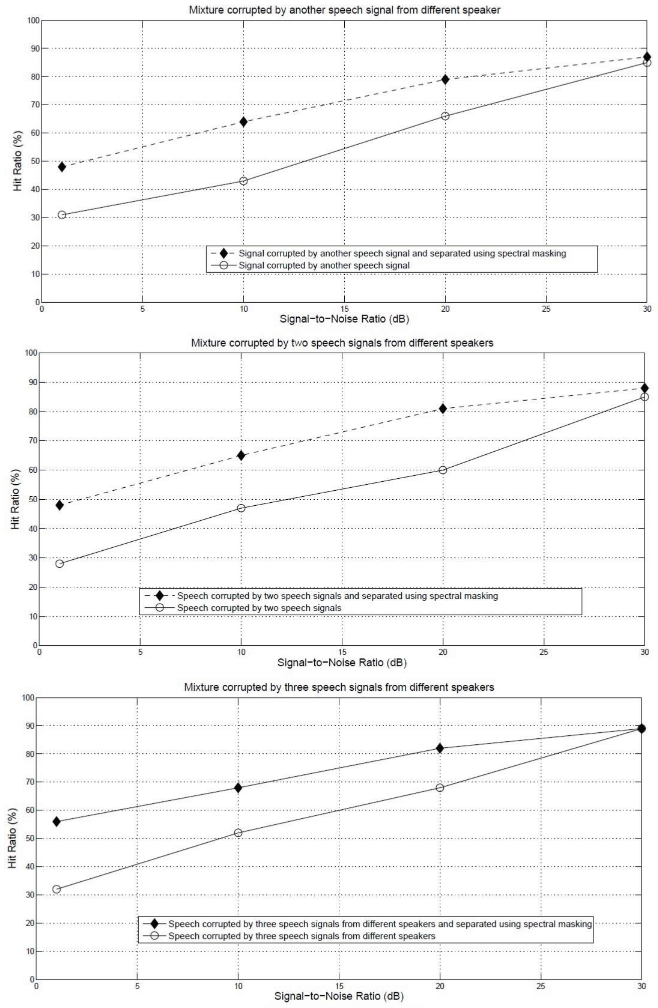

Fig. 4. Speech hit rate for the cases where the original speech signals were corrupted by other speech signals from different speakers with different signal-to-noise ratios (1, 10 and 20 dB). The limits of performance improvement of the recognition system using ideal spectral masking (with masking) were analyzed.

signal corrupted by two other speech signals from different speakers; iii) speech signal corrupted by three other speech signals from different speakers; iv) speech signal mixture corrupted by different types of noise.

According to the results shown in Fig. 4, we verify that in case the speech signal is corrupted by low level noise composed of other speech signals with a signal-to-noise ratio of approximately 30 dB, the hit rate is the same as when the ideal binary mask was applied. In some cases there was a small performance improvement when using the ideal binary mask, as shown in Tables I and II.

We also verify that in the cases where the original signals were corrupted by other speech signals with signal-to-noise ratio levels of 1, 10 and 20 dB, there is an average growth of 18 percentual points in the hit rate when applying ideal binary masking. These results show that when various signals are mixed, spectral masking technique provides a gain of approximately 10 dB in noise level attenuation, significantly improving the speech recognition systems performance, as shown in Fig. 4.

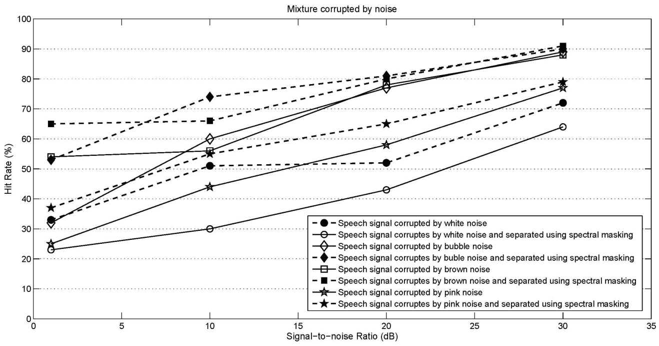

Fig. 5. Speech hit rate for the cases where the original speech signals were corrupted by different types of noise (white, pink, brown and bubble) with different signal-to-noise ratios (1, 10 and 20 dB). The limits of performance improvement of the recognition system using ideal spectral masking (with masking) were analyzed.

We show the robustness of the spectral masking techniques against four types of noise as well: white, pink, brown and bubble. In cases where the signal is corrupted by different types of noise, with SNR of approximately 30 dB (low level noise), we notice a small performance improvement of the hit rate when the ideal binary mask was applied, as shown in Table IV.

Moreover, when applying ideal binary masking with higher levels of speech human noise (1,10 and 20 dB) we observe an average growth of 18 percent in the hit rate, similar to that observed in tables I, II and III for the same levels of signal-to-noise ratio. On the other hand, applying the ideal binary masks to mixtures corrupted by white, pink and brown noise, results in an average growth of 9 % on the speech hit rate, with the same different signal-to-noise ratio.

Among the different types of noise, the experimental results reveal that the best hit rates, when applying ideal binary masking, were obtained while using the bubble noise. The worst results were obtained applying the ideal mask to white noise as in Fig. 5. This suggests that the masking spectral techniques works best for bubble noise, which is produced by human speech and justify its applications to realistic situations like human communication.

V. Conclusions

In this paper we presented an insight into the use of spectral masking techniques in time-frequency domain, as a preprocessing step for the speech signal recognition. Speech recognition systems have their performance negatively affected in noisy environments or in the presence of other voice signals. A speech recognition system based on source separation using ideal binary masks was presented in order to investigate the performance improvement in speech recognition. The signals were corrupted by noise (speech signals and different types of noise) with different signal-to-noise ratios (1, 10, 20 e 30 dB) during the tests. We show the robustness of the spectral masking techniques in the presence of four types of noise: white, pink, brown and bubble. The main contribution of this study was the analysis of the limits of performance improvement of the recognition systems using ideal spectral masking. We verified an average improvement of 18% of the hit rates for signal-to-noise ratios of 1, 10 and 20 dB. We also showed that the spectral masking techniques when applied to mixtures corrupted by other speech signals provide an average gain of 10 dB in noise level attenuation, for the same conditions of signal-to-noise ratios mentioned above. The experimental results suggest that the masking spectral techniques are more appropriate for the case when it is applied a bubble noise, which is produced by human speech, than for the case of applying white, pink and brown noise.

Acknowledgment

The authors are thankful to the Brazilian agency FAPEMIG for the financial support.

References

[1] D. Kolossa, R. F. Astudillo, E. Hoffmann and R. Orglmeister, "Independent Component Analysis and Time-Frequency Masking for Speech Recognition in Multitalker Conditions," *EURASIP Journal on Audio, Speech, and Music Processing*, 2010.

[2] T. T. Kristjansson and B. J. Frey, "Accounting for uncertainty in observations: a new paradigm for robust automatic speech recognition," in *Proceedings of the IEEE International Conference on Acustics, Speech, and Signal Processing*, 2002.

[3] V. Stouten, H. Van Hamme and P. Wambacq, "Application of minimum statistics and minima controlled recursive averaging methods to estimate a cepstral noise model for robust ASR," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2006.

[4] M. Van Segbroeck and H. Van Hamme, "Robust speech recognition using missing data techniques in the prospect domain and fuzzy masks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4393–4396.

[5] D. Kolossa, A. Klimas and R. Orglmeister, "Separation and robust recognition of noisy, convolutive speechmixtures using time-frequency masking and missing data techniques," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, vol. 13, pp. 82–85.

[6] M. *Kühne*, R. Togneri and S. Nordholm, "Time-frequency masking: linking blind source separation and robust speech recognition in Speech Recognition: Technologies and Applications," *IN-TECH, Vienna, Austria*, 2008, pp. 61–80.

[7] S. Srinivasan and D. Wang, "Transforming binary uncertainties for robust speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2130–2140, 2007.

[8] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[9] G. J. Brown and D. L. Wang, "Separation of speech by computational auditory scene analysis," in *Speech Enhancement*, J.Benesty, S. Makino, and J. Chen, Ed. Springer, New York, 2005, pp. 371–402.

[10] S. Srinivasan, N. Roman and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48,pp. 1486–1501, 2006.

[11] S. Srinivasan, N. Roman and D. L. Wang, "On binary and ratio time-frequency masks for robust speech recognition," in *Proc. International Conference on Spoken Language Processing*, 2004, pp. 2541–2544.

[12] H. Sawada, S. Araki, R. Mukai and S. Makino, "Blind extraction of dominant target sources using ICA and timefrequency masking," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2165–2173, 2006.

[13] T. S. V. Souza, G. F. Rodrigues, A. C. S. Souza, J. M. Moreira and H. C. Yehia, "Binary Spectral Masking for Speech Recognition Systems," in *Proc. 35th International Conference on Telecommunications and Signal Processing (TSP)*, 2012, pp. 432–436.

[14] E. Hoffmann, D. Kolossa and R. Orglmeister, "A batch algorithm for blind source separation of acoustic signals using ICA and time-frequency masking," in *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation*, 2007, pp. 480–487.

[15] G. Hu and D. L. Wang, "Speech segregation based on pitch tracking and amplitude modulation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 79–82.

[16] A. Jourjine, S. Rickard and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP2000)*, Jun. 2000, vol. 5, pp. 2985–2988.

[17] N. Roman, D. L. Wang and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, pp. 2236–2252, 2003.

[18] N. Roman and D. L. Wang, "Binaural sound segregation for multisource reverberant environments," in *Proc. IEEE ICASSP*, 2004, vol. 2, pp. 373–376.

[19] G. F. Rodrigues and H. C. Yehia, "Limitations of the Spectrum Masking Technique for Blind Source Separation," *Lecture Notes in Computer Science*, vol. 5441, pp. 621–628, 2009.

[20] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.*, vol. 3, no. 123, pp. 1673–1682, 2008.

[21] B. S. Kirei, M. D. Topa, I. Muresan, I. Homana and N. Toma, "Blind Source Separation for Convolutive Mixtures with Neural Networks," *Advances in Electrical and Computer Engineering*, vol. 11, no. 1, pp. 63–68, 2011. Available:http://dx.doi.org/10.4316/AECE.2011.01010

[22] A. M. Ahmad, S. Ismail and D. F. Samaon, "Recurrent neural network with backpropagation through time for speech recognition," in *Proceedings of the IEEE international symposium on communications and information technology*, 2004, vol. 1, pp. 98–102.

[23] R. P. Lippmann, "Neural network classifiers for speech recognition," *The Lincoln Laboratory Journal*, vol. 1, pp. 107–128, 1988.

[24] S. I. Amari and A. Cichocki, "Adaptive Blind Signal Processing - Neural Network Approaches," in *Proceedings of IEEE*, 1998, vol.86, no. 10.

[25] D. Kobayashi, S. Kajita, K. Takeda and F. Itakura, "Extracting speech features from human speech like noise," in *Proceedings of IEEE, Fourth International Conference on Spoken Language (ICSLP 96)*, 1996, vol. 1, pp. 418–421.

Gustavo Fernandes Rodrigues holds the degree of Electrical Engineer (CEFET-MG, 2008), Master of Electrical Engineering (UFMG, 2003) and Doctor of Electrical Engineering (UFMG, 2008). Currently, he holds the position of Professor at the Department of Telecommunications and Mechatronics Engineering of UFSJ. His research interests include signal processing, speech recognition and blind source separation.

Thiago de Souza Siqueira is an undergraduate student at the Telecommunications Engineering Program at Universidade Federal de São João del-Rei (UFSJ).

Ana Cláudia Silva de Souza holds the degree of Electrical Engineer (UFMG, 2003), Master of Electrical Engineering (UFMG, 2005) and Doctor of Electrical Engineering (UFMG, 2010). At 2008 worked at the Advanced Telecommunications Research Institute (ATR) in Japan. Currently, she holds the position of Professor at the Department of Telecommunications and Mechatronics Engineering of UFSJ. Her research interests are in Biomedical Engineering, focusing on biomedical signal processing and medical image.

Hani Camille Yehia holds the degrees of Electronics Engineer (ITA, 1988), Master of Electronic Eng. and Computer Science (ITA, 1992) and Doctor of Electrical Eng. (Nagoya University, 1997). From 1996 to 1998 he held the position of researcher at the ATR Laboratories (Japan). He holds the position of Professor at the Electronics Dept. and is currently the head of Inova-UFMG Technology Incubator. He is the head of CEFALA - Center for Research on Speech, Acoustics, Language and Music, developing research on Audiovisual production and perception of speech and Music.