SYSTEM ARCHITECTURE FOR PATTERN RECOGNITION IN ECO SYSTEMS

Benjamín Dugnol Álvarez and Carlos Fernández García

Universidad de Oviedo, Dpto. Matemáticas, Calvo Soltelo s.n., 33007 Oviedo, Spain

ABSTRACT

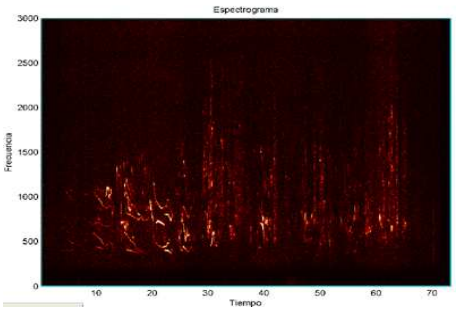
The purpose of the present work is the count and classification of the individuals in the wolf packs by processing its audio signals, supposing that we have recordings of sufficient temporary length, obtained with a single microphone. We will set out an architecture that includes the treatment of the environmental background noise, the separation of signals and its classification.

Key words: biological signals, wolf pack size, spectral subtraction, monaural signal separation, signal features, cepstral analysis, signal classification.

1. MOTIVATION OF THE PROPOSAL.

Usually wolves coexist in packs that include a pair of adults (alpha male and alpha female, leaders and ancestors) and a set of followers formed by the descendants of the last two or three years, along with some other not related component. Although the pack size is very variable, it is generally made up of a number between 2 and 10 individuals. It is usually hard to make a direct estimation of the pack size due to the difficulties of human approach to the group, by land or plane, and still preserve its integrity. Communication between the elements of the group or between the group and other near groups is made through corporal language, tracks on the ground and mainly using different sounds like: howls, barks, whispers and growls. Under ideal conditions, the howls of pack members can be heard about 15 km away. Its moderate pitch and long duration makes the sound signals have the suitable properties for transmission through forests and across tundra.

The present work has as the main purpose the counting and classification of the individuals of the pack by processing the sound signals transmition by the group, having recordings of sufficient temporary length, obtained using a single microphone.



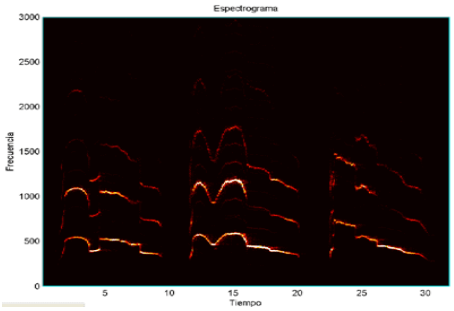
2. THE MUSIC OF THE WOLVES.

The howl of the wolves has been one of the main tools used by the biologists to asses its population. Especially, it allows the location of the territories where they move. The following studies are made using other tools.

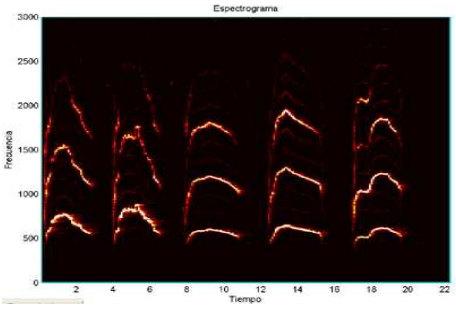
Wolves howl for very various reasons: by example, to join the dispersed pack, before and after a hunting or to communicate with another pack. The characteristics of the signal transmitted by an individual are unique, although sometimes they might be altered deliberately. This allows the components of a pack to be identified by other components of the same pack and to use the audio signal processing to

identify and to classify the individuals by the sound signals.

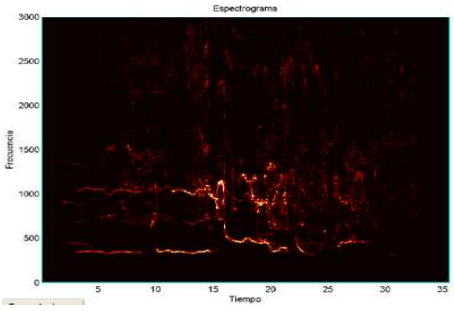
The signals describing the howl are quite simple from the point of view of the spectral characteristics. Often the pitch is constant or piecewise-constant in long temporary segments, displaying the fundamental frequency and several harmonics. Other times there is a smooth variation by pieces. The frequency modulation shows an ascending part on the onset and descendent part when each segment ends. Sometimes it shows a composition of several consecutive connected segments or with very brief interruptions.



The pups, which are individuals between 4 and 6 months, howl very often, answering any other howl they listen. They are usually made up of shorter segments than the adults segments and with bigger average fundamental frequencies.



The chorus howl follows a sequence quite easy to anticipate. One individual begins which is not necessary the alpha male. A second individual acts after one or two seconds and then the rest of the pack follows in virtually mass. The individual interruptions allow us to visually interpret some parts of the spectrogram.



As we already have suggested, it often happens that the individuals alter their voices in order to make any other near packs think that their group size is bigger then the real one. This phenomenon denominates the Effect Beau Geste and introduces an extra uncertainty in the estimation of the pack size.

3. INTRODUCTION AND STATE-OF-THE-ART.

The main mathematical methods implied in the architecture that is set out hear are:

- the separation of monaural signals
- the methods for the moderation of background noise and the detection of underlying signal (for example the Voice Activity Detector, VAD),
- the estimation of the fundamental frequency and the pitch in mixed sounds,
- the characterization of the audio signals for its classification.

A fundamental problem in signal processing is the separation of audio signals from a mixture of so. The mathematical problem of blind separation of concurrent audio signals has a degree of indetermination that we have to compensate when the number of sensors is shorter than the number of sound sources. The humans (and generally many mammals) exhibit an extraordinary capability to segregate sounds from mixtures, due to the organization of their auditory system.

Gert Cauwenberghs (Cawenberghs99) shows that a measure of the time and frequency coherence provides information to separate independent components.

Michael Zibulevsky and Barak A. Perlmutter (Zibulevsky00) suggest a process of separation in two steps: the first one consists of selecting a signal dictionary and the second one takes advantage of the sparse character of the representation of the components.

Sam T. Roweis (Roweis01) develops a method denominated re-filter that consists of recovering the sound components modulating the sub-band components of the original signal.

Guoning Hu and DeLiang Wang (Hu02) propose a system that use different mechanisms for the treatment of the LF and the HF.

Gil-Jin Jang and Te-Won Lee (Jang02) take advantage of the inherent temporary structure of the sound sources using learning sets.

Lars K. Hansen and Kaare B. Petersen (Hansen03) notice that the separation of mixtures with a single channel is a difficult problem.

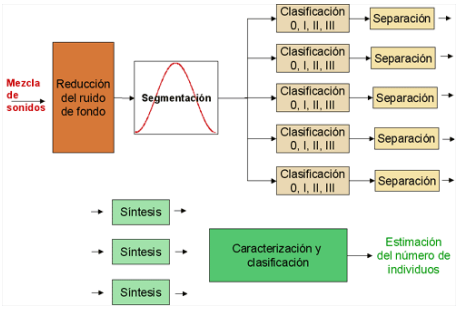
Particularly in speech signals, the spectral subtraction is one of the best addressed mechanisms for processing the background noise and for improving the signal sound. The Boll algorithm (Boll79) estimates the spectrum of the signal of underlying speech considering that the spectrum magnitudes more important than the phase from the perceptual point of view. There are several forms of the method of spectral subtraction. The greater disadvantage of these techniques consists of the introduction of the denominated musical noise (Ephraim84), which degrades the capacity of answering of the system that characterizes the signals. The approaches based on statistical models make hypothesis on the joint distribution of noise and underlying signal (Ephraim92). In the parametric techniques that use filters they suppose the signal (and possibly also the background noise) admits a autoregressive model (Lim78).

Since the 70s the efforts have been directed to the design of VAD (Voice Activity Detector) mainly based on the extraction of the energy of atoms of STFT, on measures zero-crossing rate (ZCR) or on a combination of both ((Rabiner75), (Lee86)). Recent methods have been proved that consider the regularity measures (Tucker92). J.A. Haigh and J.S. Mason (Haigh93) propose an algorithm that uses a cepstral analysis, which is characteristic of the recognition of signals algorithms.

There are many techniques of estimation multipitch . The idea of Alain de Cheveigné for a mixture of two signals is simple and theoretically perfect. It consists of minimize AMDF (Average Magnitude Difference Function) (Cheveigné91). Its disadvantage is the high computational cost. Tuomas Virtanen and Anssi Klapuri (Virtanen01) describe a method of separation of concurrent harmonic sounds. Tero Tolonen and Matti Karjalainen (Tolonen00) propose an efficient model more simple than the one proposed by Meddid and Ómard (Meddis97). Both approaches consider the human auditory system.

The investigations on the characterization of audio signals, particularly on the systems of speech recognition, have grown a lot in the two last decades. The most general context includes the recognition of the environmental sound and the effects of sound. Frank Klassner (Klassner96) showed in his thesis in 1996 a system for recognizing the environmental sounds and Alain Dufaux (Dufaux01), defended his in 2001 describing and classifying 6 sound effects. The speech recognition is possibly the best studied problem of sound characterization. Sadaoki Furui (Furui81) describes techniques for the automatic verification of a voice through the telephone using a cepstral analysis. Richard J. Mammone (Mammone96) works on the idea of identifying the inherent differences between the articulatory organs of the speech production system in order to construct a system of robust voice recognition.

4. PROPOSED ARCHITECTURE.



The recording obtained using a single microphone is digitized to a 44100 Hertz frequency. The result is a mixture of signals:

- sounds transmitted by different individuals of the group at random instants and each with undetermined length
- background noise with variable characteristics, produced by the environment that surrounds of the scene.

The first step of the process consists of a reduction of the background noise. It follows a signal segmentation with Hanning window with 50% of overlapping. The out put is a collection of signal segments classified following a content dependence criterion.

The next step consists of the separation of the signals contained in each segment already classified, that does not affect the segments without underlying signal. The information segments and the spectral data will lead to a synthesis process, considering temporal and spectral continuity criterion and also instantaneous frequencies. Thus the individual voice segments are recovered and pass to be characterize and classify.

5. PROCESSING OF THE BACKGROUND NOISE.

A classical method for reduction of background noise uses mechanisms of spectral attenuation and is based on the following hypotheses:

- The noise is random in the time domain;
- The audio signal spectrum is basically made up of elements corresponding to fundamental frequency and the sound partial;
- Due to the random nature of noise, the sound contains energy in all frequencies (broadband noise);

- The original audio signal of $x(m)$ is corrupted by an additive noise $d(m)$ uncorrelated with $x(m)$, obtaining a degraded known signal $y(m)$;
- The noise is considered stationary, at least on the blocks where the signal is subdivided.

The adopted model under these conditions is:

$$y(m) = x(m) + d(m)$$

where the signal $x(m)$ is not observable and the noise $d(m)$ can be partially observable. The purpose is estimating $\widetilde{x}$ of the original signal x based on a subset of observations y.

Considering the hypothesis on the composition of audio signals, and of fundamental and harmonic tones, the context for ideal analysis is Fourier one. Considering in addition the non-stationary of the signal, the suitable tool will be some variant of the STFT. The noise elimination technique consist of modifying the spectrum of degraded signal in each block.

Using the STFT, the spectrum of the observed data $Y(\omega)$ is obtained, where ω designates the frequency. Once obtained the spectral components, the estimation of the spectrum of the recovered signal is done by modifying the spectrum of the corrupted signal, using a function f that will be described more ahead

$$\widetilde{X}(\omega) = f(Y(\omega))$$

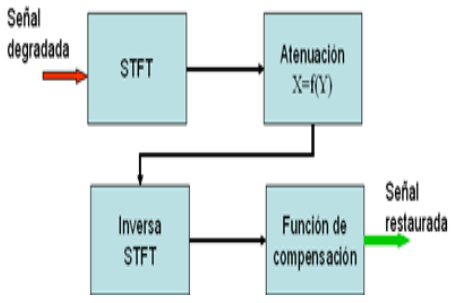
where f realizes the attenuation or spectral modification.

The modification of the signal degraded using this technique is known as spectral attenuation (STSA, Short Time Spectral Attenuation). If f is linear, the process is equivalent to the application of a real and nonnegative gain $H(\omega)$ to each frequency segment of the window spectrum observed with the purpose of estimation of the spectrum $\widetilde{X}(n,m)$ of the underlying original signal:

$$\widetilde{X}(\omega) = H(\omega) Y(\omega)$$

The specific formula for H (or equivalent for function f) is known as suppression rule and generally depends on the signal and noise power spectrum.

The spectral attenuation algorithm for eliminating the background noise is described schematically in the graph that follows:



First, the signal of audio is subdivided in blocks, applying a Hamming or Hanning window to determine STFT of the block. A level of sufficient overlapping is adopted to guarantee that the process does not introduce distortions generated when modifying the spectrum between two successive blocks. Then the function f is applied to the obtained spectrum obtaining a new version of the spectrum X, of which the inverse STFT is determined. Already in the time domain, use becomes of a function of compensation in regard to the loss of gain by effect of the window. All these operations are made in each block, so that finally the result in the advisable location is due to insert, of analogous form to since it is made actually of the standard STFT.

S. J. Boll proposes an algorithm for the suppression of acoustic noise in the speech signal. This algorithm operates a spectral subtraction by an efficient way. A segmentation of the signal by using properly overlapped Hamming windows is made to determine the spectrum of each block. This spectrum is modified by means of a filter H and, finally, the clean signal is obtained calculating the inverse transformed one and inserting each temporary block into the suitable position.

The filter of spectral subtraction H is determined replacing $|D(\omega)|$ by an amount that can be measured easily: the average value of this magnitude $\mu(\omega)$ obtained on segments of inactivity of the underlying signal. Thus,

$$\mu(\omega) = E[|D(\omega)|]$$

$$H(\omega) = 1 - \frac{\mu(\omega)}{|Y(\omega)|}$$

$$\widetilde{X}(\omega) = [Y(\omega) - \mu(\omega)] e^{j\theta_\gamma}$$

being θ_γ the phase of the spectrum of the degraded signal.

The spectral error

$$e(\omega) = \widetilde{X}(\omega) - X(\omega) = D(\omega) - \mu(\omega) e^{j\theta_\gamma}$$

it is possible to be moderated replacing $|Y(\omega)|$ by local averages:

$$\overline{|Y(\omega)|} = \frac{1}{M} \sum_{k=0}^{M} |Y_k(\omega)|$$

where subscript k makes reference to a given block of signal.

For each frequency ω satisfying

$$\frac{\mu(\omega)}{|Y(\omega)|} < 1$$

the out of the filter becomes null, so that the its new version is

$$H(\omega) = \begin{cases} 1 - \frac{\mu(\omega)}{|Y(\omega)|} & \text{if } 1 - \frac{\mu(\omega)}{|Y(\omega)|} > 0 \\ 0 & \text{if } 1 - \frac{\mu(\omega)}{|Y(\omega)|} < 0 \end{cases}$$

In absence of underlying signal, the general error $e(\omega)$ a residual noise whose presence is necessary to moderate. For it one considers that in a given block the noise fluctuates randomly in amplitude and it is possible to be suppressed replacing his current value by the minimum value obtained of adjacent blocks.

Once detected the inactivity blocks, their signal can be eliminated. In order to detect them we calculate

$$T = 20\ log_{10}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{\widetilde{X}(\omega)}{\mu(\omega)}\right| d\omega\right]$$

A block is classified like without activity if $T < T_0$ dB. The out of the filter of spectral attenuation is finally written as

$$\widetilde{X}(\omega) = \begin{cases} \widetilde{X}(\omega) & \text{if } T > T_0 \\ cY(\omega) & \text{if } T < T_0 \end{cases}$$

The algorithm can be summarized in the following steps:

- Estimation of the spectrum of the signal by using STFT,
- Estimation of the noise spectrum,
- Spectral subtraction,
- Reduction of the residual noise,
- Suppression of the noise in segments of inactivity of the underlying signal,
- Synthesis of the clean signal.

6. CLASSIFICATION OF THE SIGNAL SEGMENTS.

Once the pre-process of background noise reduction is made, the next step is the segments classification of signal in four types:

Type 0: Segments without underlying signal and with residual background noise.

Type I: Segments that possibly contain one individual voice and a residual background noise.

Type II: Segments that possibly contain a mixture of two individuals voices and a residual background noise.

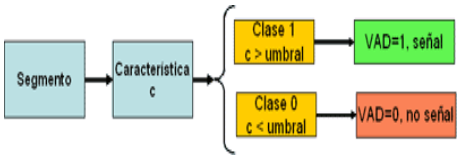
Type III: Segments that contains a mixture of three or more individuals voices.

The consideration of these types is enough to reach our goals. Statistical simulations on signals of reasonable duration (several minutes) regarding packs up to 12 individuals, illustrate the theoretical precision of the proposed architecture.

Grouping the segments in one of the defined types and their later process requires an architecture that combines different algorithms. First, a VAD (Detector of Activity of Voice) is needed to detect the segments of class 0.

VAD allows to deduce the decision mechanisms for detect if a segment of audio signal contains only noise or a mixture of noise and underlying signal. In mobile telephony they are used to identify and to compress the silence segments. The audio blocks classification in signal/no signal in noisy atmosphere is a difficult problem that it is far from a completely satisfactory solution.

Generally VAD are based on extraction of measures and characteristic magnitudes of each signal segment to compare them with a threshold. If the value of the magnitude exceeds the threshold, the segment is declared with signal (really underlying signal+background noise). On the contrary, the segment is declared of nonsignal (really background noise).



In the context of our investigation VAD robust algorithms are required for noisy conditions, and also insensitive at the level of the signal.

More formally: We classified the signal segments in one of two classes: C_0 (nonsignal) or C_1 (signal). So, the a posteriori probabilities must be compared:

$$s \in C_0 \quad if \quad p(C_0|s) > p(C_1|s) \tag{1}$$
$$s \in C_1 \quad if \quad p(C_1|s) > p(C_0|s) \tag{2}$$

Since it is difficult to calculate the a posteriori probabilities, the Bayes theorem is applied

$$p(C_i|s) = \frac{p(s|C_i)\,p(C_i)}{p(s)}$$

and the decision rule becomes

$$s \in C_0 \quad if \quad p(s|C_0)\,p(C_0) > p(s|C_1)\,p(C_1) \tag{3}$$
$$s \in C_1 \quad if \quad p(s|C_1)\,p(C_1) > p(s|C_0)\,p(C_0) \tag{4}$$

or it is declared that $s \in C_i$ being

$$C_i = argmax_C \{p(s|C)\,p(C)\} \tag{5}$$

The cepstral representation is a feature suppose the specifications of noise robustness and independence of the level. For each signal segment s, a vector is determined by the operator

$$f(s) = DFT^{-1}[log|DFT(s)|] \tag{6}$$

that calculates the iDFT of the logarithm of DFT of the segment s. The order of the cepstral analysis is identified with the number of generated coefficients.

Thus, we declare that $s \in C_i$ if

$$C_i = argmax_C \left\{ p\left(f\left(s \right) | C \right) p\left(C \right) \right\} \tag{7}$$

The practical application of the required VAD consists of comparing the cepstral vectors of the different segments with a set of training segments. Since the cepstral vectors are defined on an orthonormal representation, the comparison between vectors can be established by a weighed Euclidian distance,

$$d^{1,2} = \frac{1}{P} \sum_{k=1}^{P} \left(c^1\left(n \right) - c^2\left(n \right) \right)^2$$

being P the order of the cepstral analysis and c^1, c^2 the cepstral sequence to compare.

Once identified the Type 0 segments , the rest of the segments are classified using an multipitch estimation technique. The chosen algorithm can be described in three steps.

a) Direct use of linear model implies scale of frequencies also linear. This does not reflect properly the human hearing, since the model places poles in segments of frequency where the hearing is less sensitive. it does not arrange closely spaced samples in the low frequencies, where the hearing is more sensitive. Therefore a model of Warped Linear Prediction (WLP) is applied, that modifies the spectral resolution approximating it to the one of human the auditory system. Instead of the delay variable z^{-1} that applies a uniform resolution on the mentioned axis is used

$$D\left(z \right) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \tag{8}$$

With a sampling rate of 22 kHz we obtains the Bark scale with $\lambda = 0.63$. The synthesis filter has the form

$$A\left(z \right) = \frac{1}{1 - \sum_{k=1}^{P} a\left(k \right) D\left(z \right)^k} \tag{9}$$

where the coefficients $a(k)$ define the linear structure of the signal of audio in the customary form. The result is that it changes the timbre of the sound, doing it more intelligible, without modifying pitch of the mixture.

b) The obtained signal is processed by using a filter bank supplying two signals s_1 and s_2 on which the regularity of separated form studies first and grouped later. The detection of pitch is made with a generalized autocorrelation function by using the DFT.

For the signal $s(n)$ of length M with DFT $S(k)$ calculates the inverse transformed one of

$$\left| S\left(k \right) \right|^{\alpha} \tag{10}$$

being α a real number smaller than 2 than one adjusts of advisable form. We calculate inverse transform of

$$\left| S_1 \right|^{\alpha} + \left| S_2 \right|^{\alpha} \tag{11}$$

The determined autocorrelation function thus exhibits the points that are candidates to along with define pitch of the mixture of signals several overtones corresponding to the different components. This way, not yet it is possible to classify the segment in study.

c) The last step consists of making an improvement in the ACF before calculated. A window is applied to exhibit the positive values of the function and next the time scale is modified by a factor of 2. If this result is substracted of the windowed ACF, the overtones that corresponds to double period are eliminated. Then the fundamental frequency is shown. Clipping again to consider the positive part of the function and then expanding the time scale by a factor of 3, we eliminate the possible overtones with period 3 times the one of the fundamental frequency.

The final result shows pitch of all the signals that takes part in the mixture, so that it is possible to be decided if the segment is of type I, type II or type III.

Once classified all the segments, it is precise to determine the samples that define the beginning and the end of the nonpredominant signal. These data are indispensable for later processes in individual of the segments with two voices. With the information obtained on each segment, along with the spectral data is come to conduct an operation of synthesis, considering for it criteria of temporary, spectral continuity and of the instantaneous frequencies. Thus individual segments of voice recover for which characterization parameters are due to adopt to come finally to their classification. The number of different classes obtained is an estimation of the number of individuals of the group.

7. PROCESSING SEGMENTS WITH ONE VOICE.

We have a signal segment whose content is the sound transmitted by an only individual degraded by a background noise. Like first hypothesis, noise will be Gaussian. The case of colored noise will be examined jointly with which it corresponds to segments with two voices.

The mathematical model adopted to describe this situation is the following:

$$\begin{aligned} s\left(n \right) &= \sum_{k=1}^{P} a\left(k \right) s\left(n - k \right) + u\left(n \right) \\ y\left(n \right) &= s\left(n \right) + \nu\left(n \right) \end{aligned}$$

u and v being incorrelated Gaussian white noises, with respective variances $b = \sigma_u^2$, $d = \sigma_v^2$ and

$$\mathbf{a} = [a(P), a(P-1), ..., a(1)]$$

is the real vector corresponding to the autoregressive scheme. The same model can be expressed by a space-state formulation:

$$\begin{aligned} \mathbf{x}(n) &= F\,\mathbf{x}(n-1) + G\,u(n) \\ y(n) &= H\,x(n) + v(n) \end{aligned}$$

where F, G y H are matrices $P \times P$, $P \times 1$ and $1 \times P$ respectively:

$$F = \begin{pmatrix} 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 \\ ... & ... & ... & & ... \\ a[P] & a[P-1] & a[P-2] & & a[1] \end{pmatrix}$$

$$H = G^t = (\ 0 \quad 0 \quad ... \quad 1\ )$$

$$\mathbf{x}(n) = (\ s(n-P+1) \quad s(n-P+2) \quad ... \quad s(n)\ )^t.$$

We consider to obtain estimations of the vector $\mathbf{x}(n)$, for which we use the matrix $P \times P$:

$$Q = cov(G\,u) = \begin{pmatrix} 0 & 0 & ... & 0 \\ 0 & 0 & ... & 0 \\ ... & ... & & ... \\ 0 & 0 & & b \end{pmatrix} \tag{12}$$

and the scalar

$$R = cov(\nu) = d. \tag{13}$$

The vector $x(n)$ is constructed with the actual sample of the underlying signal and the $P-1$ samples that precede to him. By using a Kalman recursion the optimal linear estimation of $x(n)$ can be obtained known the observations

$$y^t(n) = [y(1), y(2), ..., y(n)].$$

Generally, if we have a vector of observations

$$[y(1), y(2), ..., y(m)]$$

$m < n$, we have a prediction problem. But if $m > n$, we have a smoothing problem. The estimation of $\mathbf{x}(m-L)$, $L < m$, is called fixed lag smoothing.

We suppose known all parameter of state-space model: matrices F, H y Q and the scalar d. It is designated with the symbol

$$\widetilde{x}(n|m)$$

the estimation of vector $\mathbf{x}(n)$ given $y(1), ..., y(m)$ and we define

$$\begin{aligned} e(n|m) &= \mathbf{x}(n) - \widetilde{x}(n|m) \\ P(n|m) &= E\left\{e(n|m)\,e^t(n|m)\right\} \end{aligned}$$

the estimation error and its covariance matrix. The hypothesis

$$\begin{aligned} \widetilde{x}(0|0) &= E\{\mathbf{x}(0)\} \\ P(0|0) &= E\left\{\mathbf{x}(0)\,\mathbf{x}^t(0)\right\} \end{aligned}$$

allow to dispose of initial values for Kalman recursion.

Initialize

$\widetilde{x}(0|0) = E\{\mathbf{x}(0)\}$

$P(0|0) = E\{\mathbf{x}(0)\,\mathbf{x}^t(0)\}$

Calculation

for $n = 1, 2, ...$

Temporal update: prediction

$\widetilde{x}(n|n-1) = F\,\widetilde{x}(n-1|n-1)$

$P(n|n-1) = F\,P(n-1|n-1)\ F^t + Q$

Measure update: correction

$K(n) = P(n|n-1)\ H^t\ [H\,P(n|n-1)\ H^t + R]^{-1}$

$\widetilde{x}(n|n) = \widetilde{x}(n|n-1) + K(n)\,[y(n) - H\,\widetilde{x}(n|n-1)]$

$P(n|n) = [I - K(n)\ H]\,P(n|n-1)$

In order to develop the recursion it is necessary to consider the parameters of the model. Since amount d is extracted from segments of the signal without underlying signal, it is sufficient to know the vector $\mathbf{a}$ and the scalar b :

$$\theta = [\mathbf{a}, b]$$

It is designated with $f_Y(y;\theta)$ the pdf of the random vector Y from the measures $y(1), y(2), ...$ are realizations and which depends of the parameters vector θ. The Maximum Likelihood estimation of θ it is obtained with

$$\widetilde{\theta}^{ML} = argmax_\theta \{log\, f_Y(y;\theta)\} \tag{14}$$

This problem is solved making use of an EM algorithm (expectation-maximization). The complete set of data is considered

$$z = \begin{pmatrix} y(n) \\ x_1(n) \end{pmatrix} \tag{15}$$

being

$$x_1^t(n) = [\ s(-P+1) \quad ... \quad s(n)\ ] \tag{16}$$

Let be $f_Z(z;\theta)$ the pdf corresponding to the complete vector of data.

Each iteration of EM algorithm consists of two steps:

E step (Expectation):

$$Q\left[\theta, \theta^{(l)}\right] = E_{\theta^{(l)}}\{log\, f_Z(z;\theta)\,|y\} \tag{17}$$

The next likelihood function puts together in C the terms which not depends on $\mathbf{a}$ or b:

$$\begin{aligned} &log\, f_Z(z;\theta) = \\ &= C - \tfrac{N}{2} log(b) - \tfrac{1}{2b} \textstyle\sum_{n=1}^{N} [s(n) + \mathbf{a}^t \mathbf{s}_{P-1}(n-1)]^2 \end{aligned} \tag{18}$$

where $\mathbf{s}_P$ is a column vector with $P+1$ rows:

$$s_P[n] = [s(n-P), ..., s(n)]^t \tag{19}$$

We have then

$$Q\left[\theta, \theta^{(l)}\right] = -\frac{N}{2} log(b) - \\ -\frac{1}{2b}\sum_{n=1}^{N} E_{\theta^{(l)}}\left[s(n) + \mathbf{a}^t \mathbf{s}_{P-1}(n-1)\right]^2 \tag{20}$$

Also

$$E_{\theta^{(l)}}\left(s(n) + \mathbf{a}^t \mathbf{s}_{P-1}(n-1)\right)^2 = \\ = E\left[(s(n))^2\right] + \mathbf{a}^t E\left[\mathbf{s}_{P-1}(n-1)\,\mathbf{s}_{P-1}^t(n-1)\right] + \\ +2\mathbf{a}^t E\left[\mathbf{s}_{P-1}(n-1)\,s(n)\right] \tag{21}$$

We will calculate these values in the E step of the algorithm, using the vectors which describe the actual estimation of the underlying signal.

M step (Maximization):

$$max_\theta Q\left[\theta, \theta^{(l)}\right] \longrightarrow \theta^{(l+1)} \tag{22}$$

The maximization of $Q\left[\theta, \theta^{(l)}\right]$ for θ realizes by numeric differentiation:

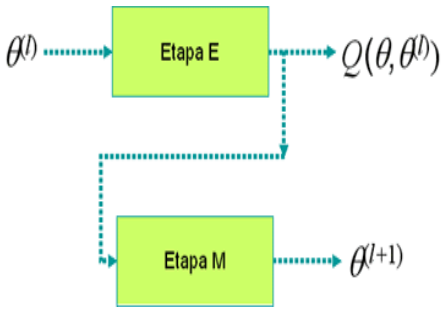
$$\frac{\partial Q\left(\theta, \theta^{(l)}\right)}{\partial \theta} = 0 \longrightarrow \theta^{(l+1)} \tag{23}$$

We deduces an algorithm that permits iterative calculation of estimation for $\mathbf{a}$ and b. For the $\mathbf{a}$ vector,

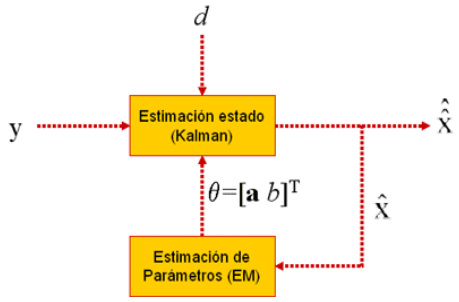
$$\mathbf{a}^{(l+1)} = \left(\sum_{n=1}^{N} E_{\theta^{(l)}}\left[\mathbf{s}_{P-1}(n-1)\,\mathbf{s}_{P-1}^t(n-1)\right]\right)^{-1} \cdot \\ \cdot \sum_{n=1}^{N} E_{\theta^{(l)}}\left[\mathbf{s}_{P-1}(n-1)\,s(n)\right] \tag{24}$$

Next, for b scalar we have:

$$b^{(l+1)} = \frac{1}{N}\sum_{n=1}^{N} E_{\theta^{(l)}}\left[s^2(n)\right] + \\ +\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{a}^{(l+1)}\right)^t E_{\theta^{(l)}}\left[\mathbf{s}_{P-1}(n-1)\,\mathbf{s}_{P-1}^t(n-1)\right] + \\ +2\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{a}^{(l+1)}\right)^t E_{\theta^{(l)}}\left[\mathbf{s}_{P-1}(n-1)\,s(n)\right] \tag{25}$$



Resuming, the signal separation or extraction process is the results of a combination of the state using a Kalman recursivity and the model parameters using the EM algorithm. The following graphic describes the process.



8. PROCESSING SEGMENTS WITH TWO VOICES.

In order to reduce the indetermination when two similar signals are mixed, we suppose that the background noise processing using spectral subtraction leads to its practical elimination. Thus, we have signal segments that contain the mixture of two voices.

We formulate the next problem: Let be s_1 and s_2 the underlying signals which we model as independent AR processes with respective parameters A_1, P_1, $\sigma_1^2 = b_1$, A_2, P_2 y $\sigma_2^2 = b_2$, and which we suppose as given. In addition we have the observations

$$y = s_1 + s_2 \tag{26}$$

for a block of length N. We will estimate s_1, so that the two processes are completely determinated. The respective pdf processes s_1, s_2 are,

$$f_{s_i}(s_i) = \frac{1}{(2\pi b_i)^{\frac{N-P_i}{2}}} exp\left(-\frac{1}{b_i} s_i^t A_i^t A_i s_i\right) \tag{27}$$

for $i = 1, 2$. They are valuable for all block samples except to the first P_i, $i = 1, 2$.

To obtain a maximum a posteriori estimate s_1^{MAP} we use the maximization of $f_{s_1|y}(s_1|y)$. Following Bayes theorem, we can write:

$$f_{s_1|y}(s_1|y) = \frac{f_{y|s_1}(y|s_1)\, f_{s_1}(s_1)}{f_y(y)}. \tag{28}$$

Besides,

$$f_{y|s_1}(y|s_1) = f_{s_1}(y - s_1) \tag{29}$$

and suppressing the scale factor $f_y(y)$

$$f_{s_1|y}(s_1|y) = \\ = \frac{exp\left[-\frac{1}{2b_2}(y-s_1)^t A_2^t A_2 (y-s_1) - \frac{1}{2\sigma_{e_1}^2} s_1^t A_1^t A_1 s_1\right]}{(2\pi b_1)^{\frac{N-P_1}{2}} (2\pi b_2)^{\frac{N-P_2}{2}}}. \tag{30}$$

Thus the MAP estimation of y_1 is

$$s_1^{MAP} = \left(\frac{A_1^t A_1}{b_1} + \frac{A_2^t A_2}{b_2}\right)^{-1} \frac{A_2^t A_2}{b_2} y \tag{31}$$

The direct algorithm for MAP estimation of s_1 implies calculation of the inverse of a large matrix, possibly ill=conditioned. A more efficient method utilizes Kalman recursions. With the notations

$$F_j = \begin{bmatrix} 0 & 1 & ... & 0 \\ 0 & 0 & ... & 0 \\ ... & ... & & ... \\ 0 & 0 & ... & 1 \\ a_j(P_j) & a_j(P_j-1) & ... & a_j(1) \end{bmatrix}$$

$$H_j = G_j^t = [\ 0 \quad 0 \quad 0 \quad ... \quad 1\ ]$$

$$\mathbf{x}_j[n] = [(\ s_j(n-P_j+1) \quad ... \quad s_j(n)\ )]$$

we can write the state equations,

$$\mathbf{x}_j(n) = F_j \mathbf{x}_j(n-1) + G_j u_j(n) \tag{32}$$

for $j = 1, 2$. A more compact model is, for both signals:

$$\begin{aligned} \mathbf{x}(n) &= F\mathbf{x}(n-1) + Gu(n) \\ y(n) &= H\mathbf{x}(n) \end{aligned}$$

being

$$\mathbf{x}(n) = \begin{bmatrix} \mathbf{x}_1(n) \\ \mathbf{x}_2(n) \end{bmatrix} \quad H = \begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \end{bmatrix}^t$$

$$F = \begin{bmatrix} F_1 & 0 \\ 0 & F_2 \end{bmatrix} \quad H = \begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \end{bmatrix}^t$$

$$u(n) = \begin{bmatrix} u_1(n) \\ u_2(n) \end{bmatrix} \quad G = \begin{bmatrix} G_1 & 0 \\ 0 & G_2 \end{bmatrix}$$

$$Q = E\left[u(n)\, u^t(n)\right] = \begin{bmatrix} b_1 & 0 \\ 0 & b_2 \end{bmatrix} \tag{33}$$

We have a linear system with no measure noise. The algorithm of Kalman recursion for perfect measure systems is the following:

Inicialize

$$\widetilde{x}(0|0) = E[(\widetilde{x}(0))]$$
$$P[0|0] = E[x(0)\, x^t(0)]$$

Calculation

for $n = 1, 2, ...$

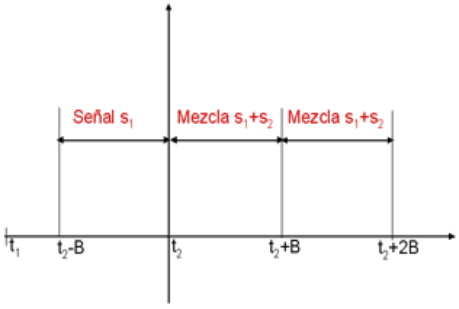
$$\widetilde{x}(n|n-1) = \widetilde{A}(n-1)\, \widetilde{x}(n-1|n-1) + K(n-1)\, y$$
$$K(n) = FP(n|n-1)\, H\left[H^t P(n|n-1)\, H\right]^{-1}$$
$$\widetilde{A}(n) = F - K(n)\, H^t$$
$$P(n|n) = \widetilde{A}(n-1)\, P(n|n-1)\, F^t + GQG^t \tag{34}$$

where $C_j = E\left[x_j(0)\, x_j^t(0)\right]$ are the covariance matrix for the AR processes, $j = 1, 2$. Considering $y = s_1 + s_2$, there is a redundance that can be used to reduce the complexity.

For the implementation of the recursion is necessary to estimate the model parameters. This estimation step will be combinated with the previous bayesian algorithm. The process utilizes an adequate window. First, we locate the initial samples of signals initiation s_1 y s_2, respectively t_1 y t_2. The work blocks are the following:

$$\begin{aligned} B_1 &= [t_2 - B, t_2] \\ B_2 &= [t_2, t_2 + B] \\ B_3 &= [t_2 + B, t_2 + 2B] \end{aligned}$$

B being the block length. For block B_1 we realizes a first estimation of coefficients a_1, which permits to make a signal s_1 extrapolation to block B_2. For B_2 we calculates then $s_2 = y - s_1$ and we applied the bayes algorithm with coefficients a_2 from s_2. The new bayesian estimated a_1, a_2 for B_2 facilitate the extrapolation to B_3, where we apply now the separation algorithm.



The problem of one voice with colored noise results as a particular case of two voices that we have described here.

9. CHARACTERIZATION, RECOGNITION AND ESTIMATION.

The audio signal characterization system adopted utilizes its production components along with properties of the human auditory system respective their perception mechanisms, that it has good characteristics of discrimination of mixtures. This system recognizes audio signals fundamentally being based on four properties of the sound: The pitch, the loudness, the duration and the timbre. The last property, the timbre or sound color, is the more multidimensional and complex. Elsewhere, the digital signal processing facilitate the characterization of all four properties with the spectrum, the pitch, the onset and offset, the amplitude envelope, dynamic attributes and the properties of transition between signal blocks.

Like already established, the pups and the adults produce various sounds: howl, barks, whispers and growls. For characterize the features of different individuals, a set of vectors based on linear models and its corresponding spectral analysis, along with time-frequency representations of the signals.

The proposed features set is the following:

A1: Linear prediction coefficients of AR model, or WLPC (warped) model,

A2: Cepstral coefficients; also with the Mel frequency scale,

A3: Delta-cepstral coefficients,

A4: Impulse response h(n) of AR model,

A5: Spectral centroid,

A6: Onset segment duration,

A7: Amplitude envelope,

A8: Amplitude modulation,

A9: Fundamental frequency,

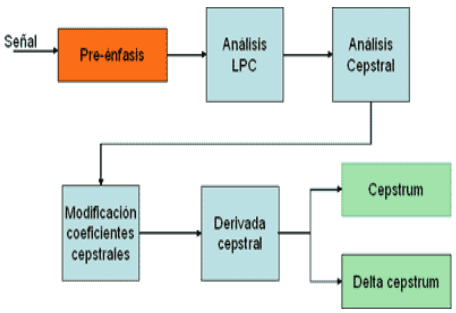
A10:1 and 2 harmonic,

A11: Frequency modulation,

A12: Spectral ratio,

A13: Normalized energy.

A scheme to calculate the cepstral coefficients is the following:



First, we normalize the signal, subtracting mean and operating an scale modification for amplitudes. Thus, the values are between -1 and 1.

We realize a pre-emphasis using a order 1 FIR with transference function

$$H(z) = 1 - \mu z^{-1} \tag{35}$$

μ being a parameter whose values are between 0.90 y 0.95. If we get

$$\mu = \frac{r(1)}{r(0)} \tag{36}$$

where $r(n)$ is the autocorrelation sequence of the input signal. The value μ is updated on each signal block.

We apply high quality windows to avoid blocking effects.

A signal $s(n)$ cepstrum is defined as the following sequence

$$c(n) = DFT^{-1}\left(log\left|DFT\left[s(n)\right]\right|\right). \tag{37}$$

According to LPC model, the signal can b e written as

$$s(n) = \sum_{k=1}^{P} a(k)\, s(n-k) + Ge(n) \tag{38}$$

where $a(k)$ are the model coefficients, $e(n)$ the excitation signal and P the order model. The a coefficients are estimated by the Levinson-Durbin algorithm.

The transference function of the model is

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^{P} a(k)\, z^{-k}} \tag{39}$$

without the gain G. Equivalently

$$\begin{aligned} H(z) = \tfrac{1}{A(z)} = \prod_{k=1}^{P} \tfrac{1}{1-z_k z^{-1}} = \\ = \sum_{k=1}^{P} \tfrac{r_k}{1-z_k z^{-1}} \end{aligned} \tag{40}$$

where r_k are the residuals and z_k the poles of $H(z)$. The poles can be written:

$$z_k = \sigma(k)\, e^{i\omega(k)} \tag{41}$$

thus $H(z)$ corresponds to a impulse response

$$h(n) = \sum_{k=1}^{P} r_k z_k^n = \sum_{k=1}^{P} r_k \sigma^n(k)\, e^{i\omega(k)n}. \tag{42}$$

The cepstral coefficient sequence can be calculated efficiently using a recursive method:

$$c(n) = -a(n) - \frac{1}{n}\sum_{k=1}^{n} kc(k)\, a(n-k) \quad n > 0 \tag{43}$$

where $a(0) = 1$ and $a(k) = 0$ when $k > P$. We get $c(0) = 0$ to avoid the gain G dependence.

The linear frequencies scale isn't the most appropriated. Then, we use Warped Linear Prediction, WLP. The $a(k)$ coefficients are calculated by the Levinson-Durbin algorithm and the cepstral sequence by the recursive formula.

The cepstral sequences based on LP or WLP describe spectral properties for a signal block, avoiding the block transitions. To solve this problem we use the cepstral derivative or delta cepstrum:

$$\frac{\partial c(n,t)}{\partial t} = \Delta c(n,t) = \mu \sum_{k=-K}^{K} c(n, t+k). \tag{44}$$

Once the cepstral analysis is cocluded, we build the vector features including the linear model coefficients (A1), the cepstral coefficients (A2), delta cepstral (A3) and the impulse response $h(n)$ (A4).

The spectral centroid (A5) is calculated for each signal window using filter banks.

$$f_o = \frac{\sum_{k=1}^{B} P_k f_k}{\sum_{k=1}^{B} P_k} \tag{45}$$

k being he filter identification whose mean square power is P_k and central frequency f_k.

The spectral ratio (A12) captures the spectrum variability segment with more elevated energy:

$$Sr\,[n] = \frac{E_n - E_{n-1}}{E_n + E_{n-1}}. \tag{46}$$

The normalized energy (A13) is determined for the frequency segment with more elevated energy

$$En\,[n] = \frac{E_n\,(f_{max})}{\sum_{k=0}^{N-1} E_n\,(f_k)} \tag{47}$$

where N is the FFT longitude.

Based on features, the signal classification have two different directions:

First, we design a two classes mechanism: adults and pups. Next we realize a initial class and another new class for each individual which not belong to an old class. The number of different classes is the pack size estimation. The number of individuals of classes pups and adults shows the pack structure.

10. CONCLUSIONS AND FUTURE WORK.

The audio separation signal using the information captured by a single microphone surrounding by background noise is a very complex problem. The proposed architecture leaves to reasonable results, although we believe that improvements could be provided for information capture and for architecture.

1. The signal capture using two microphones is very relevant. This improvement reduces the problem indetermination, make more simple the background noise processing and allow the echo signal processing.

2. The background noise processing can be realized with wavelet transform.

3. The background noise processing can be realized with Kalman recursion.

4. The separation step, the most difficult task, can be realized with although signal dictionaries, like Best Basis, Basis Pursuit or Matching Pursuit.

5. Following the human auditory system, Stéphane Maes (Maes96) suggest a method of nonlinear squeezing for to derive the amplitude and phase components of the signal and then to derive signal features: 'wastrum' instead of cepstrum.

6. The linear model selected to describe signals can be enhanced considering a new representation for excitation signal.

REFERENCES

Boll, S.F.: *Suppression of Acoustic Noise in Speech using Spectral Subtraction*. IEEE Trans. Acoust., Speech and Signal Processing, vol. 27, pp. 113-120, 1979.

Brown J. C.: *Computer Identification of Musical Instruments using Pattern Recognition with cepstral coefficients as features*, J. Acoust. Soc. Am. 105, pp. 1064-1072, 1999.

Cauweenberghs, G.: *Monaural separation of independent acoustical components*. In Proc. of IEEE Symp. Circuit&Systems, 1999.

de Cheveigné, A.: *A mixed speech F0 estimation algorithm*, ESCA (Eurospeech)-91, Genova, 445-448, 1991.

Dufaux, A.: *Detection and Recognition System for Impulsive Audio Signals*, PhD Thesis, University of Neuchâtel, IMT, April 2001.

Ephraim, Y., Malah, D.: *Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator*. IEEE Trans. Acoust., Speech and Signal Processing, ASSP-21 (6), pp. 1109-1121, 1984.

Ephraim, Y.: *A Bayesian Estimation Approach for Speech Enhancement Using Hidden Markov Models*. IEEE Transactions on Signal Processing, vol. 40, 4, 1992.

Eronen, A.: *Automatic Musical Instrument Recognition*. Master of Science Thesis, Department of Information Technology, Tampere University of Technology, 2001.

Furui, S.: *Cepstral Analysis Technique for Automatic Speaker Verification*, IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 29, No. 2, 1981, pp. 254-272.

Gannot, S.: *Algorithms for single microphone speech enhancement*, M.Sc. thesis, Tel-Aviv Univ., Apr. 1995.

Godsill S.J.,Tan,CH.: *Removal of low frequency transient noise from old recordings using model-based signal separation techniques*. In Proc. IEEE Workshop on Audio and Acoustics, Mohonk, NY State, Mohonk, NY State, October 1997.

Haigh J.A., Mason, J.S.: *Robust Voice Activity Detection using Cepstral Features*. In IEEE TENCON, pp. 321-324, China, 1993.

Hansen, L.K., Petersen, K.B.: *Single Channel Source Separaton is Hard*. Conference Proceedings of ICA 2003.

Härmä, A.: *Implementation of frequency-warped recursive filters*, Signal Processing, 80 (3), pp. 543-548, February 2000.

Harrington, F. H.: *Chorus howling by wolves: Acoustic structure, pack size, and the beau geste effect,*Bioacoustics 2(2), 117-136, 1989.

Harrington, F. H.: *What's in a Howl?*Nova online. http://www.pbs.org, November 2000.

Hu, G., Wang, D.: *Monaural Speech Segregation Based on Pitch Tracking and Amplitude Modulation*, Proc. of IEEE Int. Conf. on Acoustic, Speech ans Signal Processing, 2002.

Jang G.-J., Lee T.W.: *A probabilistic Approach to Single Channel Blind Signal Separation*. In Advances in Neural Information Processing Systems 15 (NIPS 02), 2002.

Karjalainen, m.: *Auditory Interpretation and Application of Warped Linear Prediction*, Consistent & Reliable Acoustic Cues for sound analysis. Aalborg, Denmark, Sunday September 2nd 2001.

Kil D.H., Shin, F.B: *Pattern Recognition and Prediction with Aplications to Signal Characterization*. American Institute of Physics, 1996.

Klassner, F.: *Data Reprocessing in Signal Understanding Systems*. Ph. D. thesis, Department of Computer Science, University of Massachusetts Amherst, September 1996.

Lee, H.H., Un C.K.: *A Study of On-Off characteristics of convesational Speech*. IEEE Transactions on Communications, Vol. COM-34, 6, pp. 630, 1986.

Lim J.S., Oppenheim A.V.: *All-pole modelling of degraded speech.*IEEE Trans. Acoust., Speech and Signal Processing, ASSP-21 (6), pp. 1109-1121, 1984.

Daubechies I., Maes S.: *Nonlinear Squeezing of the Continuous Wavelet Transform Based on Auditory Nerve Models*. In Wavelet in Medicine and Biology, edited by Aldroubi and Unser, CRC Press, 1996.

Mammone, R. J., Zhang, X. and Ramachandran R. P.: *Robust Speaker Recognition: A Feature-Based Approach*, IEEE Signal Processing Magazine, Vol. 13, No. 5, Sept. 1996.

Meddis, R., O'Mard, L.: A unitary model for pitch perception. J. Acoust. Soc. Amer., Vol. 102, pp. 1811-1820, sept. 1997.

Popescu, D.C. Zeljkovic, I.: *Kalman Filtering of Colored Noise for Speech Enhancement*, Proceedings 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP' 98, vol. 2, pp. 997-1000, May 12-15 1998, Seattle, Washington.

Rabiner, L.R., Sambur, M.R.: *An algorithm for determinig thr endpoint of isolated utterances*, The Bell System Technical Journal, Vol. 54, Num 2, pp. 297, 1975.

Roweis, T.: *One microphone source separation*. In Advances in Neural Information Processing Systems 13 (NIPS 00), 2001.

Tolonen, T., Karjalainen, M.: *A Computationaly Efficient Multipitch Analysis Model*. IEEE Transactions on Speech and Audio Processing, Vol. 8, No. 6, november 2000.

Tooze, Z.J., Harrington, F.H., & Fentress, J.C.: *Individually distinct vocalizations in timber wolves (Canis lupus).*Animal Behaviour, 40, 723-730., 1990.

Tucker R.: *Voice Activity Detection using a periodicity measure*, IEE Proceedings, Vol. 139, 4, 1992.

Virtanen, T., Klapuri, A.: *Separation of Harmonic Sounds Using Multipitch Analysis and Iterative Parameter Estimation*. Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, 2001.

Zibulevsky, M., Pearlmutter, B.A.: *Blind Source Separation by Sparse Decomposition in a Signal Dictionary*, 2000.