Non-speech Sound Feature Extraction Based on Model Identification for Robot Navigation

Yolanda Bolea[1], Antoni Grau[1], and Alberto Sanfeliu[2]

[1]Automatic Control Dept, Technical University of Catalonia UPC,
[2]Robotics Institute, IRI/CSIC, UPC
08028-Barcelona, Spain
{yolanda.bolea;antoni.grau}@upc.es, sanfeliu@iri.upc.es

Abstract. Non-speech audio gives important information from the environment that can be used in robot navigation altogether with other sensor information. In this article we propose a new methodology to study non-speech audio signals with pattern recognition techniques in order to help a mobile robot to self-localize in space domain. The feature space will be built with the more relevant coefficients of signal identification after a wavelet transformation preprocessing step given the non-stationary property of this kind of signals.

1 Introduction

Sound offers advantages for information systems in delivery of alerts, duration information, encoding of rapidly incoming information, representing position in 3-D space around the person and her localization. Hearing is one of human beings most important senses. After vision, it is the sense most used to gather information about our environment. Despite this, little research has been done into the use of sound by a computer to study its environment. The research that has been done focuses mainly on speech recognition [1], [2], while research into other types of sound recognition has being neglected. In robotics, non-speech audio has been ignored in front of artificial vision, laser beams and mechanical wave sensors beyond the audible spectrum. But the study and modeling of non-speech audio can help greatly to robot navigation and localization in the space domain. The existing research in non-speech sound is incipient and focuses on signal processing techniques for feature extraction with the use of neural networks as a classification technique [3], [4]. In this article a new technique based on pattern recognition techniques in order to locate a robot in the space domain by non-speech audio signals is proposed. The feature space will be built with the coefficients of model identification of audio signals. Due to their non-stationary property wavelet decomposition is needed as a preprocessing step. We also propose a technique (transform function) to convert the samples in the feature space into the space domain, based in the sound derivative partial equation described in [1]. In section 2 the feature selection and feature vector are described as soon as the procedure to obtain the transform function. In section 3 we present an experiment in order to test the proposed algorithms and techniques.

2 Non-speech Audio Feature Extraction Approach for Localization in Space Domain

In this section we propose a new localization in space domain approach from non-speech audio signals that will be applied on a robot in an industrial environment, the approach follows the next steps: 1) measurement and data preprocessing. 2) MAX models signals identification by the wavelet transform; 3) feature selection, feature extraction and its correspondence with the space domain. Non-speech audio signal generated by any audio source (industrial machinery, appliances, etc.) is continuous by its nature. Preliminary, non-speech signal preprocessing includes sampling the analog audio signal with a specific frequency and to convert it into a discrete set of samples. Sampling interval should be chosen in such a way that essential information be preserved. In this case, due to the audio signal form we have followed the same criteria as [5] in order to choose the sampling frequency because its similarity to speech signals.

2.1 Model Identification by the Wavelet Transform and Feature Selection

Non-speech audio signal have the property of non-stationary signal in the same way that many real signals encountered in speech processing, image processing, ECG analysis, communications, control and seismology. To represent the behavior of a stationary process is common the use of models (AR, ARX, ARMA, ARMAX, OE, etc.) obtained from the experimental identification [6]. The coefficient estimation can be done with different criteria: LSE, MLE, among others. But in the case of non-stationary signals the classical identification theory and its results are not suitable. Many authors have proposed different approaches to modeling this kind of non-stationary signals, that can be classified: i) assuming that a non stationary process is locally stationary in a finite time interval so that various recursive estimation techniques (RLS, PLR, RIV, etc.) can be applied [6]; ii) a state space modeling and a Kalman filtering; iii) expanding each time-varying parameter coefficients onto a set of basis sequences [7]; and iv) nonparametric approaches for non-stationary spectrum estimation such a local evolving spectrum, STFT and WVD are also developed to characterize non-stationary signals [8].

To overcome the drawbacks of the identification algorithms, wavelets could be considered for time varying model identification. The distinct feature of a wavelet is its multiresolution characteristic that is very suitable for non-stationary signal processing [9]. Wavelet transform can decompose $\boldsymbol{L}^2(R)$ space to a linear combination of a set of orthogonal subspace adaptively which divide the whole frequency bands into a series of subbands from high to low frequency, representing the multiresolution characteristics of the original signal.

As non-speech audio signals are non-stationary and have very complex waveforms because of the composition of various frequency components, a signal transformation is performed. The idea of signal transformation is to separate the incoming signal into frequency bands. This task may be solved with the use of filter bank or wavelet transform, as psychoacoustics has associated human hearing to non-uniform critic bands. These bands can be realized roughly as a four-level dyadic tree. For sampling at 8kHz

the frequencies of the dyadic tree are 0-250Hz, 250-500Hz, 500-1000Hz, 1000-2000Hz, 1000-2000Hz and 2000-4000Hz. Each input signal are decomposed in 4 levels, that is, the audio signal $S_i = A4_i + D4_i + D3_i + D2_i + D1_i$, where $A4_i$ is the approximation of the original S_i signal and Dj_i (j=1,4) are the detail signals for S_i.

The wavelet transform have been done with the Daubechies wavelet, because it captures very well the characteristics and information of the non-speech audio signals. This set of wavelets has been extensively used since its coefficients capture the maximum amount of the signal energy [9].

A MAX model (Moving Averaging Exogenous) represents the sampled signals in different points of the space domain because the signals are correlated. We use the closest signal to the audio source as signal input for the model. Only the model coefficients need to be stored to compare and to discriminate the different audio signals. This would not happen if the signal were represented by a AR model because the coefficients depend on the signal itself and, with a different signal in every point in the space domain, these coefficients would not be significative enough to discriminate the audio signals. When the model identification is obtained by wavelets transform, the coefficients that do not give information enough for the model are ignored. The eigenvalues of the covariance matrix are analyzed and we reject those coefficients that do not have discriminatory power. For the estimation of each signal the approximation signal and its significative details are used following the next process: i) model structure selection; ii) model parameters calibration with a estimation model (the LSE method can be used for its simplicity and, furthermore a good identified model coefficients convergence is assured); iii) validation of the model.

Let us consider the following TV-MAX model and be $S_i = y(n)$,

$$y(n) = \sum_{k=0}^{q} b(n;k)u(n-k) + \sum_{k=0}^{r} c(n;k)e(n-k) \tag{1}$$

where $y(n)$ is the system output, $u(n)$ is the observable input, which is assumed as the closest signal to the audio source, and $e(n)$ is a noise signal. The second term is necessary whenever the measurement noise is colored and needs further modeling. In discrete time, wavelet expansions are computed through filter banks. Now we expand the coefficients $b(n;k)$ and $c(n;k)$ onto a wavelet basis,

$$y(n) = T_1(n) + T_2(n) \qquad \text{where} \tag{2}$$

$$T_1(n) = \sum_{k=0}^{q} \sum_{m} \zeta_{Jmax,m}^{(b_k)} \left[\tilde{h}_0^{(Jmax)} (n - 2^{Jmax} m) u(n-k) \right] + \sum_{k=0}^{q} \sum_{j=Jmin}^{Jmax} \xi_{j,m}^{(b_k)} \left[\tilde{h}_1^{(j)} (n - 2^{j} m) u(n-k) \right] \tag{3}$$

$$T_2(n) = \sum_{k=0}^{r} \sum_{m} \zeta_{Jmax,m}^{(c_k)} \left[\tilde{h}_0^{(Jmax)} (n - 2^{Jmax} m) e(n-k) \right] + \sum_{k=0}^{r} \sum_{j=Jmin}^{Jmax} \xi_{j,m}^{(c_k)} \left[\tilde{h}_1^{(j)} (n - 2^{j} m) e(n-k) \right] \tag{4}$$

Let $h_0(n)$ and $h_1(n)$, be a dyadic Perfect Reconstruction Filter Bank (PRFB).

Then, for a fixed k, the wavelet coefficients, corresponding to the low-resolution and the detail signal of $b(n;k)$, are given by

$$\zeta_{l,m}^{(b_k)} = \sum_{l} h_0(l) b(2m - l; k) \qquad \text{and} \tag{5}$$

$$\xi_{l,m}^{(b_k)} = \sum_l h_1(l) b(2m-l;k) \tag{6}$$

respectively. Therefore the signal $b(n;k)$ can be reconstructed from $\zeta_{l,m}^{(b_k)}$ and $\xi_{l,m}^{(b_k)}$ by the synthesis equation,

$$b(n;k) = \sum_m \zeta_{l,m}^{(b_k)} \tilde{h}_0(n-2m) + \sum_m \xi_{l,m}^{(b_k)} \tilde{h}_1(n-2m) \tag{7}$$

where $\tilde{h}_l(n) = h_l(-n),\ l = 1,2$. See reference [9] for further details. In order to obtain the $c(n;k)$ coefficients we follow the same procedure.

2.2 Feature Extraction and Spatial Recognition

The coefficients for the different models will be used as the feature vector, which can be defined as X_S, where

$$X_S = (b_1, b_2, \ldots^{q+1)} \ldots, c_1, c_2, \ldots^{r+1)} \ldots) \tag{8}$$

where $q+1$ and $r+1$ are the amount of **b** and **c** coefficients respectively. From every input signal a new feature vector is obtained representing a new point in the $(q+r+2)$-dimensional feature space, *fs*. For feature selection, it is not necessary to apply any statistical test to verify that each component of the vector has enough discriminatory power because this step has been already done in the wavelet transform preprocessing.

This feature space will be used to classify the different audio signals entering the system. For these reason we need some labeled samples with their precise position in the space domain. (In the following section an specific experiment is shown). When an unlabeled sample enters the feature space, the minimum distance to a labeled sample is computed and this measure of distance will be used to estimate the distance to the same sample in the space domain. For this reason we need a transformation function which converts the distance in the feature space in the distance in the space domain, $f_T : \Re \rightarrow \Re$, ($f_T$: ((*q*+*r*+2)-D *fs*) $\rightarrow$ (2-D x-y space domain), note that the distance is an scalar value, independently of the dimension of the space where it has been computed.

The Euclidean distance is used, and the distance between to samples S_i and S_j in the feature space is defined as

$$d_{fs}(S_i, S_j) = \sqrt{\sum_{k=0}^{q} (b_{kS_i} - b_{kS_j})^2 + \sum_{k=0}^{r} (c_{kS_i} - c_{kS_j})^2} \tag{9}$$

where b_{kSi} and c_{kSi} are the **b** and **c** coefficients, respectively, of the wavelet transform for the S_i signal. It is not necessary to normalize the coefficients before the distance calculation because they are already normalized intrinsically by the wavelet transformation.

This distance computation between the unlabeled sample and labeled samples is repeated for the three closest samples to the unlabeled one. Applying then the transformation function f_T three distances in the x-y domain are obtained. These distances indicate where the unlabeled sample is located. Now, with a simple process of geometry, the position of the unlabeled sample can be estimated. The intersection of the three

circles, ideally yields a unique point, corresponding to the position of the unlabeled sample. In the practice, the three circles intersection yields an area proportional to the error of the whole system. The position of the sample is approximated by the centroid of this area.

$$f_T : d_{fs}(S_i,S_k) \rightarrow d_{xy}(S_i,S_k)=r_i;\ f_T : d_{fs}(S_j,S_k) \rightarrow d_{xy}(S_j,S_k)=r_j;\ f_T : d_{fs}(S_p,S_k) \rightarrow d_{xy}(S_p,S_k)=r_p$$

where S_i, S_j and S_p are three labeled samples and r_i, r_j and r_p are the distances in the space domain to the unlabeled sample S_k. The distance is understood as a radius because the angle is unknown.

Because there exist the same relative distances between signals with different models, and with the knowledge that the greater the distortion the farther the signal is from the audio source, we choose those correspondences (d_{xy},d_{fs}) between the samples that are closest to the audio source equidistant in the d_{xy} axis. These points will serve to estimate a curve of n-order, that is, the transformation function f_T. Normally this function is a polynomial of 4th order and there are several solutions for a unique distance in the feature space, that is, it yields different distances in the x-y space domain. We solve this drawback adding a new variable: previous position of the robot. If we have an approximate position of the robot, its speed and the computation time between feature extraction samples, we will have a coarse approximation of the new robot position, coarse enough to discriminate among the solutions of the 4th-order polynomial. In the experiments section a waveform for the f_T function can be seen, and it follows the model from the sound derivative partial equation proposed in [1].

In the figure 1 the localization system can be shown, including the wavelet transformation block, the modeling blocks, the feature space and the spatial recognition block which has as input the environment of the robot and the function f_T.

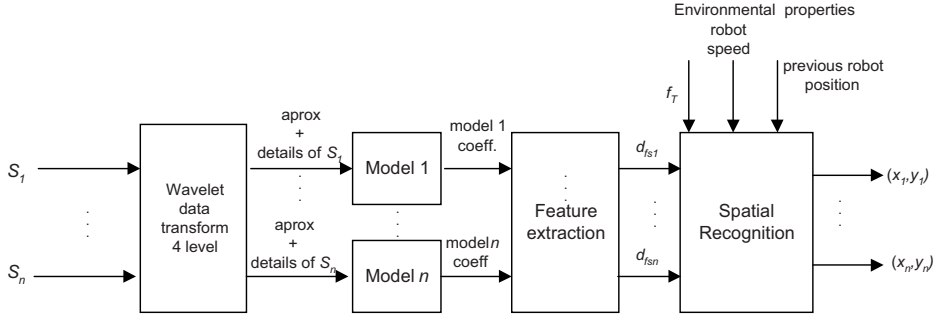

Fig. 1. Localization system in space domain from non-speech audio signals.

3 Experimental Results

In order to prepare a setting as real as possible, we have used a workshop with a CNC milling machine as non-speech audio source. The room has a dimension of 7 meters by 10 meters and we obtain 9 labeled samples (from S_1 to S_9), acquired at regular positions, covering all the representative workshop surface. With the dimensions of the room, these 9 samples are enough because there is not a significative variance when oversampling. In figure 3 (right) the arrangement of the labeled samples can be observed. The robot [10] enters the room, describes a predefined trajectory and gets off. In its trajectory the robot picks four unlabeled samples (audio signals) that will be used as data test for our algorithms (S_{10}, S_{11}, S_{12} and S_{13}). The sample frequency is 8kHz and a capacitive microphone is used.

First, in order to obtain the 9 models coefficients corresponding to the 9 labeled non-stationary audio signals, these signals are decomposed by the wavelet transform in 4 levels, with one approximation signal and 4 detail signals, figure 2. For the whole samples, the relevance of every signal is analyzed. We observe the more significative decomposition to formulate the prediction model, that is, those details containing the more energy of the signal. With the approximation ($A4_i$) and the detail signal of 4th level ($D4_i$) is enough to represent the original signal, because the mean and deviation for the $D3_i$, $D2_i$ and $D1_i$ detail signals are two orders of magnitude below $A4_i$ and $D4_i$. Figure 2 (up right) shows the difference between the original signal and the estimated signal with $A4_i$ and $D4_i$. Practically there is no error when overlapped. In this experiment we have chosen the Daubechies 45 wavelet transform because it yields good results, after testing different Daubechies wavelets.

After a initial step for selecting the model structure, it is determined that the order of the model has to be 20 (10 for the $A4_i$ and 10 for $D4_i$ coefficients), and a MAX model has been selected, for the reasons explained above.When those 9 models are calibrated, they are validated with the error criteria of FPE (Function Prediction Error) and MSE (Mean Square Error), yielding values about 10^{-6} and 5% respectively using 5000 data for identification and 1000 for validation. Besides, for the whole estimated models the residuals autocorrelation and cross-correlation between the inputs and residuals are uncorrelated, indicating the goodness of the models.

These coefficients form the feature space, where the relative distances among all the samples are calculated and related in the way explained in section 2 in order to obtain the transform function f_T. With these relations, the curve appearing in figure 3 (left) is obtained, under the minimum square error criteria, approximated by a 4th-order polynomial with the following expression:

$$f_T = d_{fs} = -9.65e(-10)d_{xy}^4 + 1.61e(-5)d_{xy}^3 - 8.49e(-2)d_{xy}^2 + 144.89d_{xy} + 107.84 \quad (10)$$

which is related with the solution of the sound equation in [1] with a physical meaning.

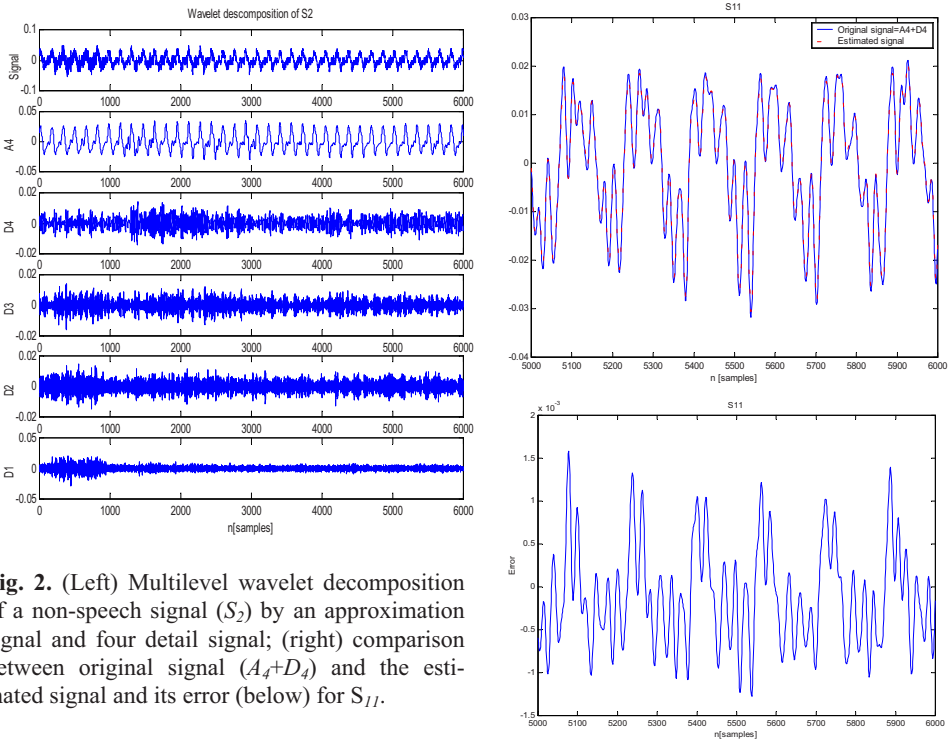

Fig. 2. (Left) Multilevel wavelet decomposition of a non-speech signal (S_2) by an approximation signal and four detail signal; (right) comparison between original signal (A_4+D_4) and the estimated signal and its error (below) for S_{11}.

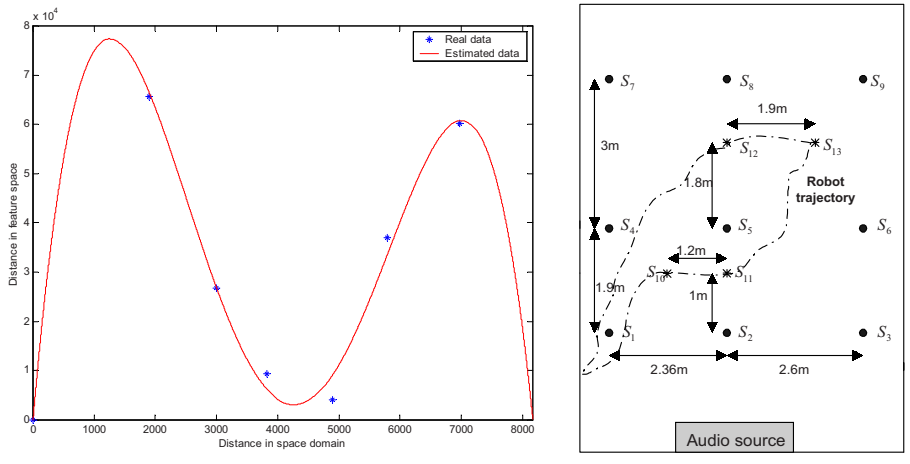

Fig. 3. (Left) Transform function f_T; (right) robot environment: labeled audio signals and actual robot trajectory with unlabeled signals (S_{10}, S_{11}, S_{12}, S_{13}).

With the transform function f_T we proceed to find the three minimum distances in the feature space to each unlabeled sample respect the labeled ones, that is, for audio

signals S_{10}, S_{11}, S_{12} and S_{13}, respect S_1, ..., S_9. We obtain four solutions for each signal because each distance in the feature space crosses four times the f_T curve. In order to discard the false solutions we use the previous position information of the robot, that is the $(x_i,y_i)_{prev}$ point. We also know the robot speed (v = 15cm/sec) and the computation time between each new position given by the system, which is close to 3 sec. If we consider the movement of the robot at constant speed, the new position will be $(x_i,y_i)_{prev}$ ± (450,450)mm. With this information we choose the solution that best fits with the crossing circles solution. In table 1, the recognition rate for each estimated position in space domain are presented, in any case there is an error bigger than the 15%, and in one case the error is under the 0.5%.

Table 1. Rate of spatial recognition results for unlabeled samples respect their actual position.

Original signal	S_{10}		S_{11}		S_{12}		S_{13}	
Cartesian coord.	x_{10}	y_{10}	x_{11}	y_{11}	x_{12}	y_{12}	x_{13}	y_{13}
Recognition rate (%)	90.4	85	97.98	87.69	89.18	99.58	88.35	94.42

4 Conclusions

With the methodology presented in this article we have achieved some interesting results that encourage the authors to keep on walking in this research field. The introduction of more that one audio source is also a new challenge. The experimental results show a narrow correspondence with the sound physical model and this demonstrates a high reliability of the proposed methodology.

References

[1] Rabenstein, R., Trautmann, L.: Digital Sound Synthesis by Physical Modeling. *Symposium on Image and Signal Processing an Analysis* (ISPA'01), 2001.

[2] Y. Zhong et al, "Binary Quantization of Feature Vectors for Robust Text-Independent Speaker Identification", *IEEE Trans on Speech and Audio Proc*, vol. 7, no 1, Jan. 1999.

[3] Spevak, C. & Polfreman, R. (2001). Distance measures for sound similarity based on auditory representations and dynamic time warping. In: *Proceedings of the InternationalSymposium on Systematic and Comparative Musicology*, pp. 165-170. Jyväskylä, Finland, 2001.

[4] Tzanetakis, G. & Cook, P. (2000). *MARSYAS: A framework for audio analysis*. Organised Sound 4(3), Cambridge University Press.

[5] Bielińska, E.: *Speaker identification*. Artificial Inteligence Methods, AI-METH 2002.

[6] Ljung, L.: *System identification: Theory for the user*. Prentice-Hall, 1987.

[7] Charbonnier, R., Barlaud, M., Alengrin, G., Menez, J.*:* Results on AR-modeling of nonstationary signals. *IEEE Trans. Signal Processing* 12 (2) (1987) 143-151.

[8] Kayhan, A.S., Ei-Jaroudi, A., Chaparro, L.F.: Evolutionary periodogram for nonstationary signals. *IEEE Trans. Signal Process.* 42(6) (1994) 1527-1536.

[9] Tsatsanis, M.K., Giannakis, G.B.: Time-varying system identification and model validation using wavelets. *IEEE Trans. Signal Process.* 41(12) (1993) 3512-3523.

[10] Sanfeliu, A. et al.: MARCO: A mobile robot with learning capabilities to perceive and interact with its environment", *IX Symposium on Pattern Recognition and Image Processing*, (SNRFAI'01), pp. 219-224, Castellón, Spain, 2001.