Internal Use Only（非公開）

TR－SLT－0044

A High Resolution Time Delay Estimation Using Interchannel Linear Prediction

チャネル間線形予測による高分解能時間差推定

Sebastien GASNIER
セバスチャン　ガスニエ

August 30, 2003

ABSTRACT:

This document will present the researches I conducted in ATR during my five months internship. A Direction of Arrival (DOA) Estimator was studied. To estimate sound source locations or the directions of arrival (DOA) with precision, a high-resolution time-delay estimator which can handle fractional sample delay is needed. I then studied an inter-channel, linear prediction based time-delay estimator using a microphone array. The resultant linear prediction filter is an interpolator estimated by minimizing the inter-channel prediction error squares.

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目２番地２ TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288, Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

LIST OF APPENDIX

ASJ: Acoustic Society of Japan.

ASR: Automatic Speech recognition.

ATR: Advanced Telecommunications Research Institute International

CCG: Cross-Correlogram

CLIPS: Communication Langagiere et Interaction Personne-systeme (communication through language and human-machine interaction): French laboratory based in Grenoble and working in collaboration with ATR for the "C-Star project phase III" on the research topic of Automatic Speech Translation.

ENSERG: Ecole Nationale Superieure d'Electronique et de Radioelectricite de Grenoble.

ICASSP: International Conference on Acoustics, Speech and Signal Processing: This is a major conference on speech processing. It is held every year and groups researchers from all over the world.

ICSLP: International Conference on Spoken Language processing: it's another major speech processing conference where many researchers present their state of the art achievements and work.

LP: Linear Prediction.

SLT: Spoken Language Translation Research Laboratories.

SIR: Signal to Interference Ratio

SNR: Signal to Noise Ratio

ACKNOWLEDGEMENT

I would like to thank **Dr. Satoshi Nakamura** (ATR-SLT Dept1 head), head of SLT dept 1 and **Dr. Mitsunori Mizumachi** (ATR-SLT PhD Researcher), my supervisor, for their scientific and technical support as well as for their warm welcome in ATR and in Japan.

I would also like to thank **Dr. Frank K. Soong** (ATR-SLT Invited Researcher) for his very helpful advice and the supervising of my work.

Thank you also to **Yukiko Ishikawa** and **Makiko Tatsumi** (ATR – SHIEN, foreigners support group) for their kindness and their advices about daily life in Japan, travels, …

Thank you to all the members of SLT - Department 1 for their warm welcome and especially to **Toshiharu Horiuchi**, **Toshiki Endo** and **Shigeki Matsuda** for the nice moments spent together.

Thanks to **Laurent Besacier** (CLIPS, France, 'Maitre de conference') for his advises and support from France.

Thanks to **Mathieu, Nicolas, Christian**, **Ayako**, **Chiyako**, **Amori** and **Yuichiro** for making my stay in Japan so enjoyable.

This research was supported in part by the Telecommunications Advancement Organization in Japan entitled, "A Study of speech dialogue translation technology based on a larger corpus".

ABSTRACT

This document will present the researches I conducted in ATR during my five months internship. A Direction of Arrival (DOA) Estimator was studied. First, I studied a To estimate sound source locations or the directions of arrival (DOA) with precision, a high-resolution time-delay estimator which can handle fractional sample delay is needed. I then studied an inter-channel, linear prediction based time-delay estimator using a microphone array. The resultant linear prediction filter is an interpolator estimated by minimizing the inter-channel prediction error squares.

INTRODUCTION

State-of-the-art speech recognition systems are known to perform reasonably well when the speech signals are captured in a noise-free environment using a close-talking microphone. Many commercial applications have spread in the market for variety of uses. The demand for such systems would increase with the decrease of the error rate of these systems. However, many of the applications in this field do not take place in noise-free environments. Furthermore, a close-talking microphone is often not suitable in many of the possible applications. So as the distance between the speech source and the microphone gets longer, the signal is increasingly corrupted by background noises and reverberation effects so that the recognition accuracy. And for some situations the location can't be changed easily such as in the case of meeting rooms.

An interesting solution for this problem is the use of multiple microphones to capture the speech signal. Microphone arrays record the speech signal simultaneously over a number of spatially separated channels. Many array-signal-processing techniques have been developed to combine the signals in the array to achieve a substantial improvement in the signal-to-noise ratio (SNR) of the output signal.

Currently, microphone array-based speech recognition is performed in two independent stages: array processing and recognition. Array-processing algorithms, typically designed for speech enhancement, process the captured waveforms and the output waveforms are passed to the speech recognition system.

The goal of this work is to focus on the speech enhancement part of the array-processing algorithms. The speech enhancement part can itself be divided in two different stages: the localization of the source and then the application of the beamforming algorithms. In this review, we'll just concentrate on the localization algorithms of the speech source. Actually, the more accurate the localization finder algorithm will be, the better performance for the beamforming algorithms we can expect.

In microphone array signal processing, a high resolution DOA estimation is often desired for directional beamforming. The temporal resolution of a beamformer is, in general, limited by the sampling frequency : the coarse sampling interval of a low sampling frequency may not be adequate to resolve a fractional sample time delay [1]. As a result, traditional methods of DOA estimates usually use higher sampling frequency (i.e. higher than the Nyquist rate) to improve the time delay estimates (e.g. Cross-correlogram technique [2] & [3]) or they need some extra prerequisite such as the number of sources for achieving a higher temporal resolution (e.g., the MUSIC algorithm [4]).

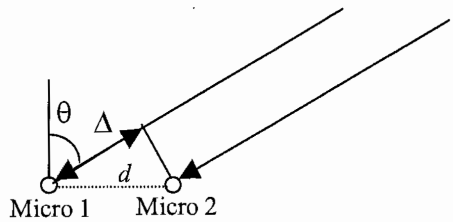


Figure 1: Representation of the problem (2 microphones case)

Under a far-field assumption, the incident signals from a sound source are considered as plane waves and the signal samples at the two microphones are:

$$\begin{cases} x_1(t) = s(t) + n_1(t) \\ x_2(t) = s(t-\delta) + n_2(t) \end{cases}$$

where $x_1(t)$ and $x_2(t)$ are the two received microphone signal samples from the sound source, $s(t)$, at time t , the arrival time difference between them is δ and two independent measurement noise samples, $n_1(t)$ and $n_2(t)$, are linearly added to the received source signals. Here, we simplify the problem by assuming that there is no attenuation or multi-path transmissions between the source and the two microphones but in general, the following proposed algorithm can still be applied but with different estimation accuracy if the environment is highly reverberant.

We want to estimate the delay, which in turn is used for estimating the DOA of the sound source, θ, which is given as:

$$\theta = \arcsin\left(\frac{c\cdot\delta}{d}\right)$$

where c and d are the sound velocity and the spacing between the two microphones. According to the above expression the estimate of θ is non-linearly dependent on the time delay estimate through the arcsin function.

OBJECTIVE

My objective as I presented this in the previous part is to estimate the Direction of Arrival (DoA) of the speech signal with precision. The goal is to obtain enough accuracy to apply beamforming techniques afterwards. As I explained in the previous paragraph, the sampling rate is very important for the accuracy of the sampling rate. However, the computation time and the real time needed in many ASR applications imply the use of lower sampling rate.

In this study, the objective is to find an high resolution DoA estimator based on low frequency sampling rate. Thus, a Linear Prediction based algorithm was studied. The experiments conducted aimed to compare the new algorithm with a baseline one (the Cross-Correlogram) and then to test this algorithm in different SNR conditions : simulated and real conditions. A comparative study of the different parameters influent in the algorithm was also carried out such as the LP filter length, the parameters influencing the accuracy of the estimation.

Generalities

Consider the signals picked up by two microphones *i* and *j*, which belong to a microphone array. Let these signals be named $x_1(n)$ and $x_2(n)$. The cross-correlation between the two signals is defined as :

$$R_{x_i x_j} = E[x_i(n-\tau)x_j(n)]$$

The operator *E[f(n)]* is called the expected value of *f(n)* and for an observation window of *N* samples of *f(n)* , an estimate of the expected value can be written as

$$\hat{E}[f(n)] = \frac{1}{N}\sum_{i=1}^{N} f(i)$$

Consider the signals arriving at two microphones with a delay of □ samples between them

$$\begin{cases} x_1(k) = s(k) + n_1(k) \\ x_2(k) = s(k - \square) + n_2(k) \end{cases}$$

The term $n_i(k)$ represents noise at the microphones. The Fourier transforms of these microphone signals can be expressed as

$$\begin{cases} X_1(w) = S(w) + N_1(w) \\ X_2(w) = S(w)e^{-jw\delta} + N_2(w) \end{cases}$$

Thus the cross power spectral density with no pre-filtering is given by

$$\begin{aligned} \Phi_{x_1 x_2}(w) &= X_1(w)X_2^*(w) \\ &= S(w)S^*(w)e^{-jw\delta} + S(w)N_2^*(w) + S^*(w)N_1(w)e^{-jw\delta} + N_1(w)N_2(w) \end{aligned}$$

The term $S(w)S^*(w)$ in the previous equation can be replaced by Φ_{SS} which represents the power spectral density of the source signal. Thus we have :

$$\Phi_{x_1 x_2}(w) = \Phi_{SS}(w)e^{-jw\delta} + S(w)N_2^*(w) + S^*(w)N_1(w)e^{-jw\delta} + N_1(w)N_2(w)$$

Since the noise signals are assumed to be uncorrelated with the signal and each other, the last three terms do not contribute towards the cross-correlation computation, which is an inverse Fourier transform operation.

$$R_{x_x x_2}(\tau) = F^{-1}[\Phi_{SS}(w)e^{-jw\delta}]$$
$$= R_{ss}(\tau - \delta)$$

Thus we find that the cross-correlation function is just the incident signal's autocorrelation centered at delay □. Ideally we would have liked the cross-correlation function to be a delta function at □ so that the peak can be easily picked out. In reality, the auto-correlation of the incident signal ends up spreading the cross-correlation function around the delay □.

$$R_{x_x x_2}(\tau) = R_{ss}(\tau) * \sum_i \alpha_i \delta(t - \delta_i)$$
$$= \sum_i R_{SS}(\tau - \delta_i)$$

This will result in peaks in the cross-correlation function at each delay. If the delays are very close to each other and if the spread in the signal auto-correlation is very large, then the peaks may merge thus decreasing the resolution of the estimate.

Implementation

I've implemented a Cross-Correlogram (CCG) based DoA estimator on Matlab environment. At first, I used a sub-band Cross-Correlogram based estimator.
The diagram flow of the program is as follow :

Variables Initialisation
⇨ distance between the microphones, sampling frequency, number of microphones, …

Loading signal samples

Artificially delayed the sample on the different channels

Cross-Correlogram Process

- **Gammatone Filterbanks**
- **Delaying the channels samples**
- **CCG computing and plotting**
- **Peaks Finder and ITD estimate**

DoA Estimation

Figure 2: Diagram Flow of the CCG algorithm

Below, an example of the kind of Graph, we can obtain with the previous program:

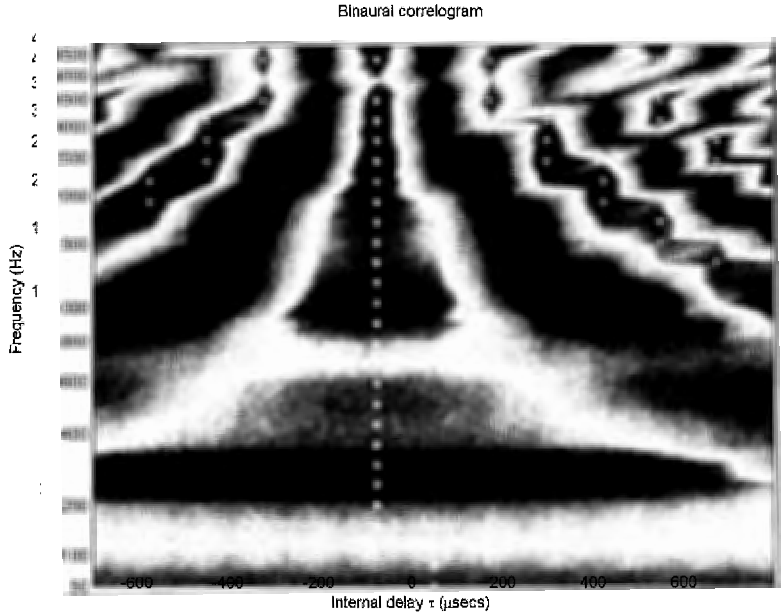


Figure 3: 2D representation of CCG segmented in sub-bands

The white stars on the previous figure show the global maxima in each subbands, wheras the white crosses show the local maxima, The alignment of the maximas show the probable DoA of the speech source on the previous figure. Below the same kind of figure on a 3D plot

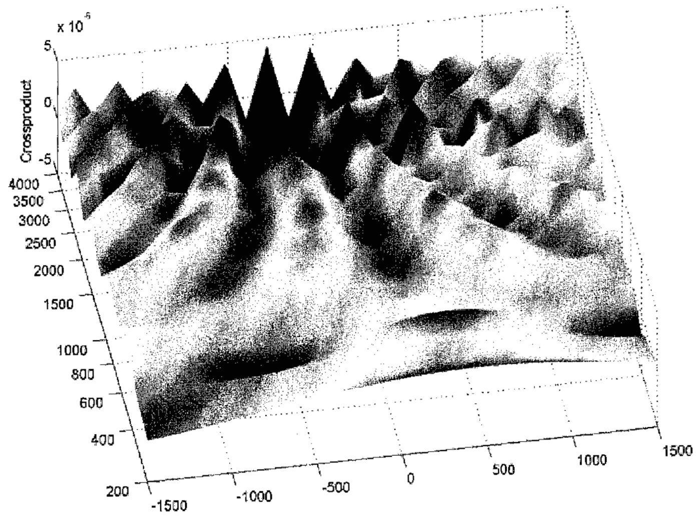


Figure 4: 3D representation of CCG segmented in sub-bands

According to the previous diagram flow, the cross-correlation product we'll be computed in each sub-bands in order to find the cross-correlation peaks in each subband and its related DoA.
In the following of this report, our baseline is the previous cross-correlation techniques in full band.

Generalities:

Linear Prediction (LP) is proposed here to estimate the time delay. A block of received samples, typically in a 30 ms window, is used for formulating the prediction equations of LP.

Considering an N-sample window, the matrix equation is used to predict the signal $\vec{x}_1$ of the first channel, from x_2 of the second channel or vice versa, as:

$$\underbrace{\begin{bmatrix} x_2(\Delta) & \cdots & \cdots & \cdots & x_2(-\Delta) \\ x_2(\Delta+1) & \cdots & \cdots & \cdots & x_2(-\Delta+1) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_2(N+\Delta) & \cdots & \cdots & \cdots & x_2(N-\Delta) \end{bmatrix}}_{\mathrm{X}_2} \cdot \underbrace{\begin{bmatrix} h_{-\Delta} \\ \cdots \\ h_0 \\ \cdots \\ h_\Delta \end{bmatrix}}_{\vec{h}} \cong \underbrace{\begin{bmatrix} x_1(0) \\ \cdots \\ \cdots \\ \cdots \\ x_1(N) \end{bmatrix}}_{\vec{x}_1}$$

The block data matrix X_2 is Toeplitz. The product $\mathrm{X}_2\vec{h}$ is an estimate of $\vec{x}_1$ with a filter $\vec{h}$, which is an "interpolation" or "smoothing" filter of length $(2\Delta+1)$. If we define the error vector $\vec{e}$ as:

$$\vec{e} = \vec{x}_1 - \mathrm{X}_2\vec{h}$$

its power is then:

$$\vec{J} = \vec{e}^{\,T}\vec{e}$$

where ^{T} is the transpose operator.

The Least Squares (LS) solution is obtained by setting the gradient of J with respect to $\vec{h}$ to zero and the "normal equation" is then solved as :

$$\frac{\partial\vec{J}}{\partial\vec{h}} = 0 \Rightarrow \vec{h} = (\mathrm{X}_2^T\mathrm{X}_2)^{-1}\cdot(\mathrm{X}_2^T\vec{x}_1)$$

The length parameter Δ is chosen to be long enough to cover the full range of possible time delay between two channels, i.e.,

$$\Delta \geq \frac{d\cdot f_s}{c}$$

where f_s is the sampling frequency.

The LP parameters are then processed through an interpolation function for example, a sinc function, to find the estimated time delay which can be of non-integer sampling periods between the two microphones. The delay which yields the largest magnitude of the "continuous" interpolated filter is our estimated time delay. An edge effect usually happens for large DOA angles, where the coefficients can become highly asymmetrical, leading to larger errors in the interpolated filter for estimating the maximum magnitude. The problem can be alleviated by repositioning the LP equations to make the “interpolator” essentially symmetrical again.

Implementation:

I’ve implemented the previous algorithm on Matlab.
The diagram Flow of the algorithm is quite similar to those from the CCG algorithm.

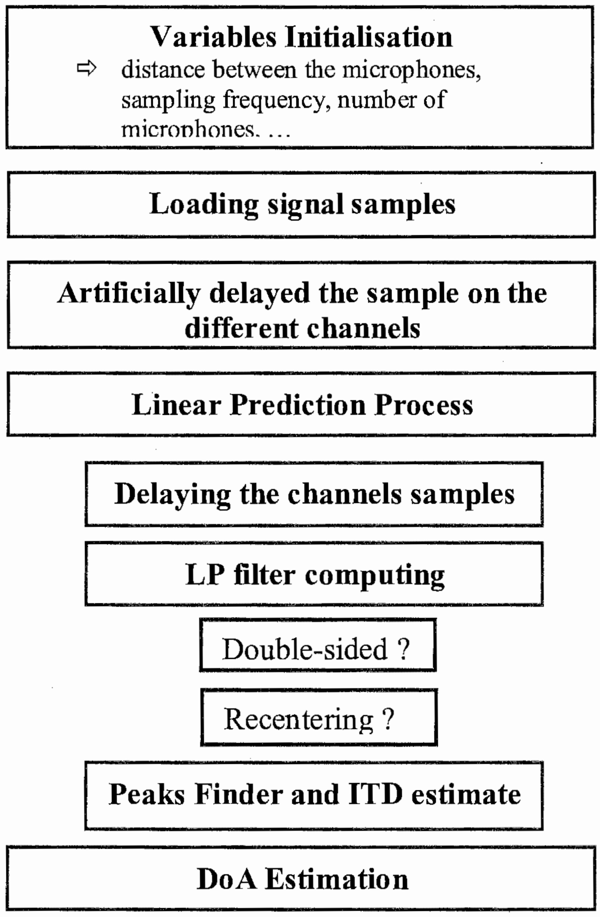


Figure 5: Diagram Flow of the LP algorithm

Simulation Experiments

Experimental Setup

Speech utterances of connected digit strings in the Aurora 2 database [5] were used in our experiments. Assuming the incoming speech signals are plane waves, we simulated the signals received by the two microphones using computer. The simulated data was then used to estimate the time delay. In the two first experiments, two independent machine generated white noises were added digitally at the two microphones to the speech signals. Also, different noise samples in Aurora 2 database including: street noise, train noise, airport noise, etc. were used.

Below a representation of the waveform of a connected digits :

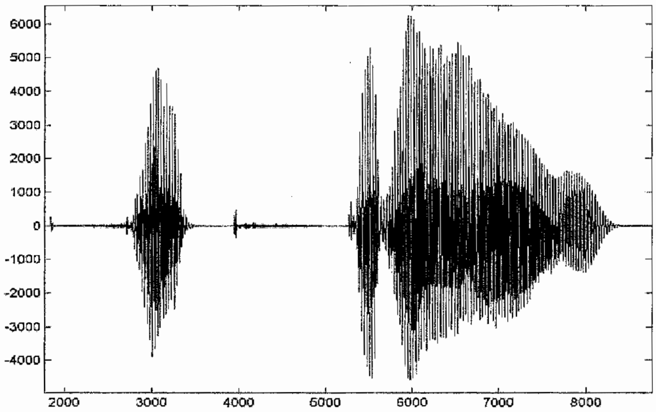


Figure 6: Connected digits Waveform

Performance of the Linear Prediction

Speech samples of ten digits were used in our first experiment where only background (silence) frames were edited out. Noises were added to the speech signals at the two receiving microphones. They were machine generated, independent random Gaussian samples. The SNR was set at 20dB. A LP filter of 13 coefficients (Δ=6) was used. The baseline algorithm for comparison is the cross-correlogram technique [2-3].

The results of the first experiment are shown in *Fig.1*. The performance of the baseline is much poorer than that of the proposed LP approach. This lack of precision of the baseline, especially at higher DOA angles, is due to the low sampling frequency used (8kHz).

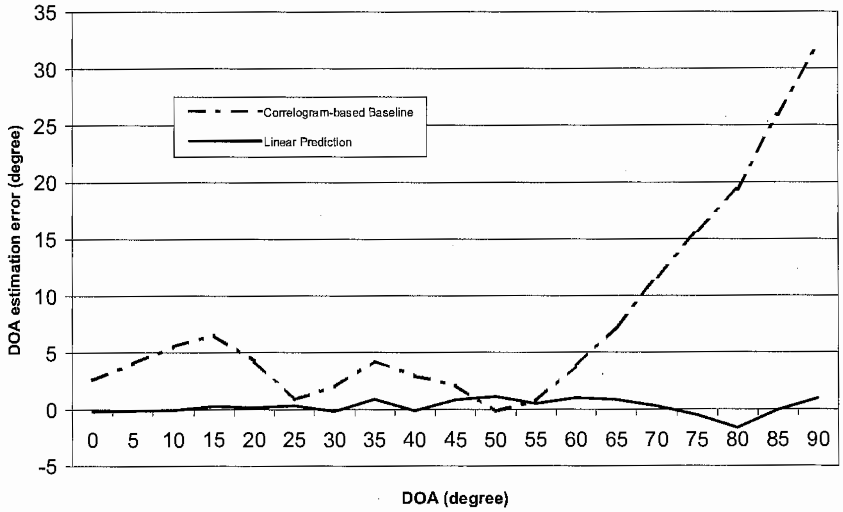


Figure 7: LP filter vs CCG performances

Effects of the LP order (2Δ+1)

In this experiment, independent white noises on each channel were used to estimate the influence of the filter length parameter Δ in the algorithm. In *Fig. 2*, filters with different lengths from 13 to 29, were tested. The higher Δ is, the lower the LP error power and the DOA estimation error is.

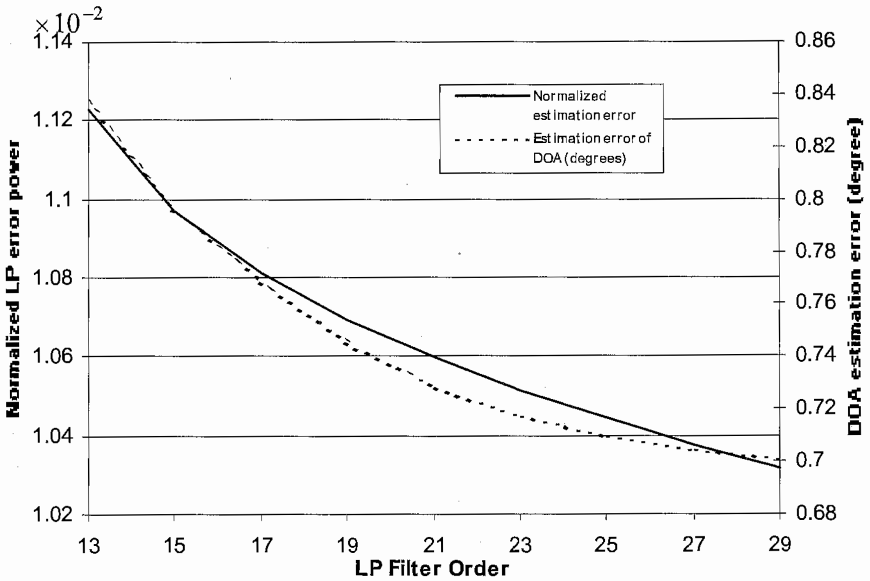


Figure 8: LP filter order influence

In the following experiments, the LP filter order chosen will be equal to 13.

3.3 Robustness of the LP method

In the third experiment, the robustness of the algorithm against different types of noise as interference was tested at various Signal-to-Interference Ratios (SIRs). Real noise samples from the Aurora 2 database are used. At different SIRs from 20dB to 0dB were used in this experiment. The DOA of the noise interference is fixed at 50 degrees from the center to the left (counterclockwise) whereas the speech source's DOA varies from 0 to 90 degrees from the center to the right (clockwise). The results of this experiment are shown in *Fig.3*. At high (20dB) to low (10dB) SIRs, the LP algorithm performs very well. But at an even lower SIR of 5dB, larger DOA errors start to appear.

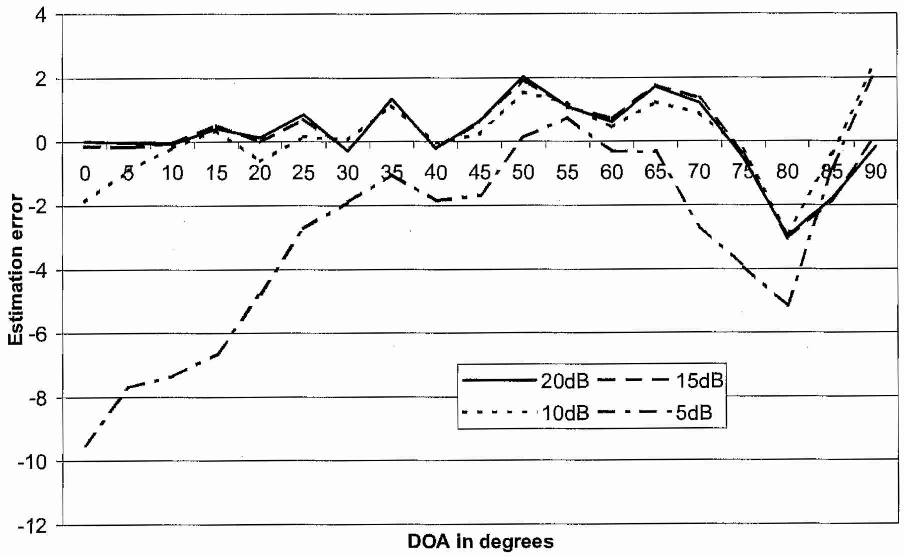


Figure 9: Robustness of LP algorithm

Real Environment Experiments:

Experimental Setup

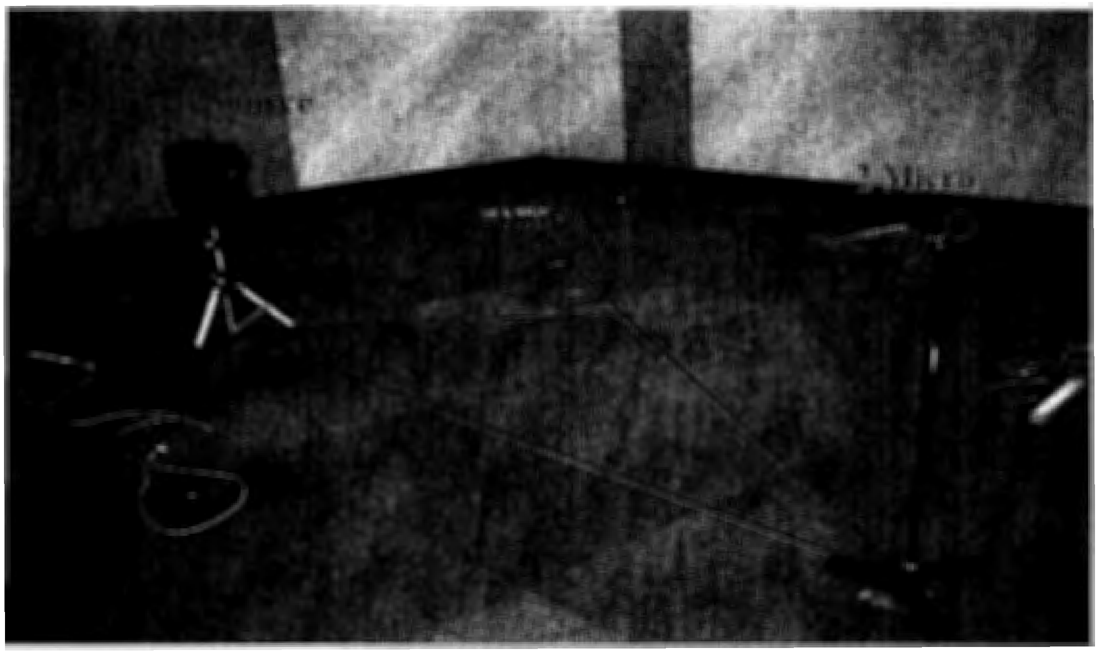


Figure 10: Experimental setup for real environment experiment

The objective of this experiment was to test the algorithm in a real environment room with reverberation and some background noises (air conditioner noise) and also to get the impulse response of the room in order to prepare others simulations results with different reverberation time. In order to realize these experiments, we used a DAT player/recorder, a 2-microphone array with a spacing equal to 20cm between the two microphones. The sampling rate for the recording and the playback is 48kHz.

So, on the figure below, we can see the signal used as desired signal:

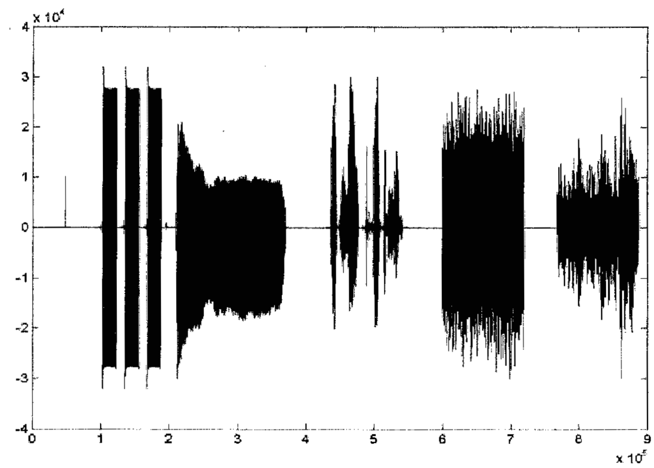


Figure 11: Playback signal for real environment experiment

The first part of the signal is a variable sine wave signal

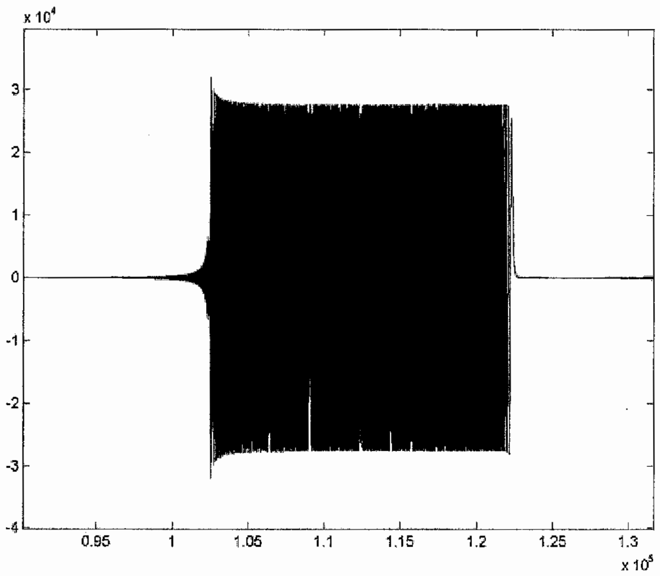


Figure 12: Variable Sine Wave signal

The method for computing the impulse response of the room is given in Annexe 2

The other parts of the signal used are respectively a self-recorded vowel “ehhhhh”, connected digits from Aurora 2 database, airport noise from Aurora 2 database and a machine generated white noise.

Unfortunately, the low value of the SNR during this recording session, the data were processed through an high pass filter before being processed through the Linear Prediction Algorithm : only the frequencies above 200 Hz are kept.

Experimental Results

On the figure below, are the results of the experiments with a DoA of 0 degree (Speaker in Front of the 2 microphones)

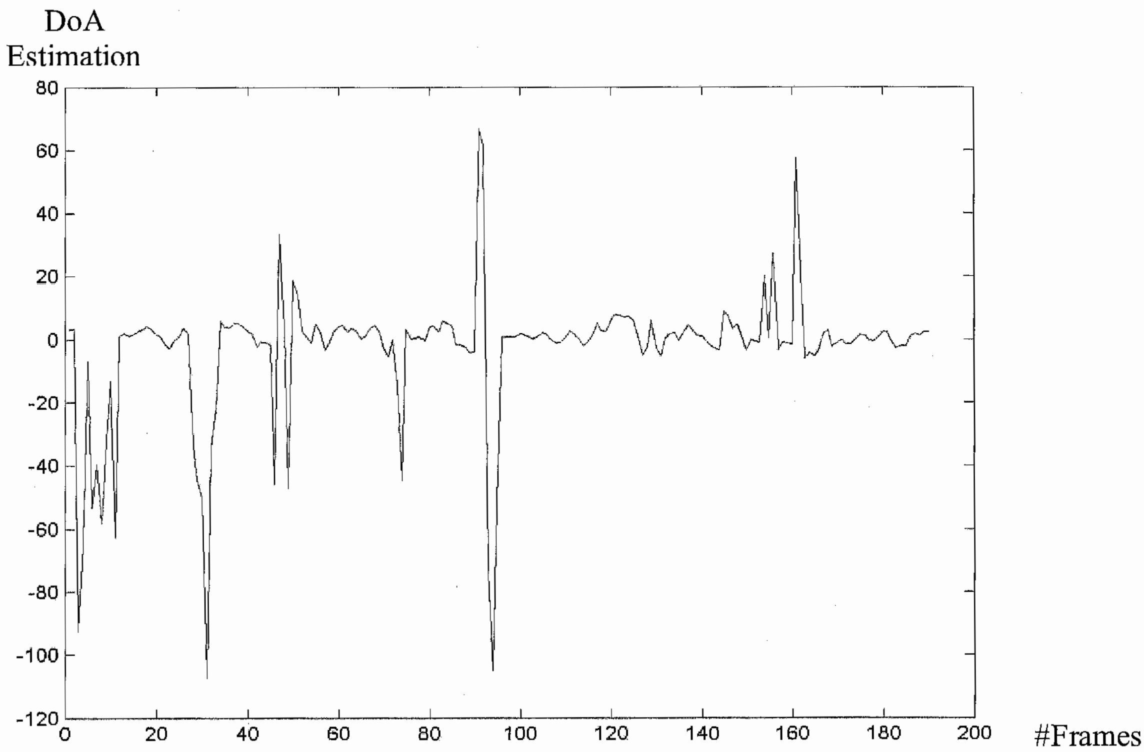


Figure 13: DoA estimation in real condition with a 0degree DoA

This results chow a quite good accuracy of the LP algorithm over the frames. Some frames are corrupted with correlated noises, which are the reasons why there are this quit important error in the estimation of the direction of arrival.

On the next figure, we can observe the evolution of the coefficients of the LP filter over the time (at each 30ms windows) for a 30degrees direction of Arrival.

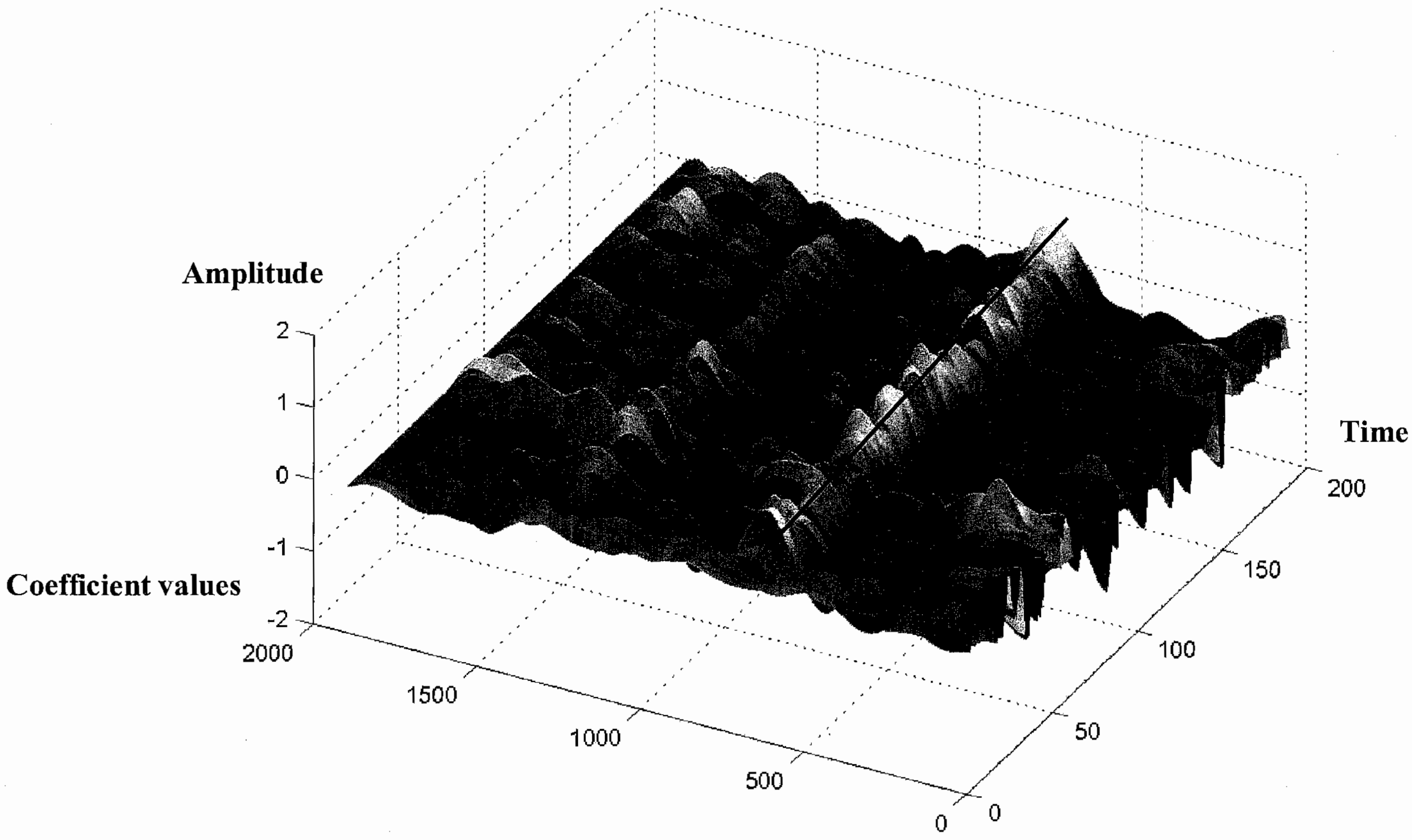


Figure 14: Evolution of the coefficients values over the frames

This figure just above just shows the evolution of the values of the LP filter's coefficients. We can see clearly that the desired source with the biggest peaks. But there are also some peaks with high values for other coefficients than the desired ones. These kind of secondary peaks are responsible for the error in the DoA estimation seen on the previous figure (*figure 13*).

In this section, I will stress the points that could be improved if further work was done on this topic.

As it has been shown in the previous paragraphs, the Linear Prediction estimation offer good accuracy results for DoA and even for low SNR. However, there are still some features for the LP algorithm to be studied further :

- The tracking of a moving, i.e, computing the DOA of the speech source, which is moving over the time.
- The accuracy of the LP algorithm in reverberant conditions (partly checked with the experiment in real conditions)
- The case of Multi-microphones (>2), the number of microphones should increase the performance of the LP algorithm by reducing the DOA estimation error.

Another enhancement of this algorithm would be the use of a LP-based DOA with robust voice-activity detector (VAD) : This kind of system would allow to track only the speech source DOA and not those from the different background noises as seen in the previous experiments.

Finally, another idea to be studied is the relatively little changes in the position of the source over the different measurement frames. The ideas would be then to use this little changes as a constraint.

GLOBAL CONCLUSION: OVERALL REVIEW

As a global conclusion, I will that the work achieved was quite efficient since in comparison to the CCG based estimator, the LP based estimator showed good performances.
Several features of this estimator were highlighted :

- The importance of the LP filter order: the longer the order is, the more accurate the predictor is.
- Robusts estimates are obtained even in low SNR cases.

This internship was my first experience in the field of research. It was a really exciting experience : I could meet and talk with researchers from all over the world, exchanging ideas. It was a very pleasant thing for me to work in ATR and many people helped me during my stay. I hope I can come again to Japan and to ATR in the future.

[1] T.I. Laakso, V. Valimaki, M. Karjalainen, and U. Laine, “Splitting unit delay” , IEEE signal processing Mag. pp. 30-60, Jan. 1996.

[2] K. J. Palomäki, G. J. Brown and D. L. Wang, “A binaural auditory model for missing data speech recognition in noisy and reverberant conditions”, Proc. CRAC - Eurospeech’01, satellite workshop, 2001.

[3] L.R. Bernstein and C. Trahiotis, “The normalized correlation: Accounting for binaural detection across center frequency”, J. Acous. Soc. Am., pp. 100, 1996.

[4] S.U. Pillai, Array Signal Processing, pp. 8-107 1989

[5] H. G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions”, Proc. ISCA ITRW ASR2000 "ASR: Challenges for the Next Millennium," 2000.

[6] R.E. Blahut, Fast Algorithms for Digital Signal Processing, Addison-Wesley, 1985.

APPENDIX

Appendix 1: Reverberation Time Measurement

Reverberation Time Measurements

A Little Theory

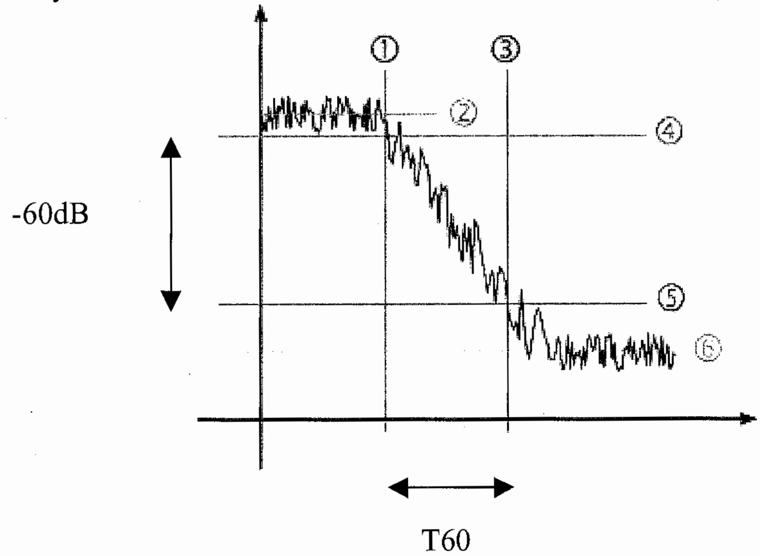
The reverberation time is defined as the time it takes for the sound level to decay by 60dB after a sound source has been switched off.
Rooms with small amounts of absorbing materials will have longer reverberation times than rooms with more absorbers.
For normal rooms, with reasonable amounts of absorbers, the reverberation time is given by:

$$\mathbf{T=0.16V/A}$$

in which **T** is the reverberation time in seconds; **V** is the room volume in cubic metres; **A** is the absorption of the room, measured in equivalent square metres open window (i.e. if the absorber has half the absorbing property as an open window, it should be used with half its surface area in the formula) and finally **0.16** is an empirical constant determined by Wallace C. Sabine and published in 1898.
This model cannot be used in rooms with excessive amounts of absorbers, such as anechoic chambers.

Measuring the Reverberation Time

To measure the reverberation time you will need a sound source and instrumentation able to capture the sound decay. Theoretically, you have two options with respect to the sound source; impulse excitation or noise excitation. However, the ISO 140 Series of Standards require the use of noise excitation. It is important that the noise is broadbanded enough to cover the entire frequency range of interest.
Although the reverberation time is defined as the time it takes for the sound to decay 60dB, this is seldom possible to measure. Imagine a room with a background noise level of 55dB. To be able to measure a 60dB decay the noise source acoustical output level must then be typical more than 120dB (55 + 60 + another 5dB distance to the background noise level), which is a requirement very hard to meet for practical systems. The frequency range required for field measurements is 100-3150Hz, but many measurements are now made up to 5000Hz which is the requirement for laboratory measurements. Optionally, you may extend the frequency range downwards to 50Hz, which would make the requirement even tougher with respect to output levels.
Instead of measuring the complete decay, we normally measure a 15, 20, 30 or 40dB decay which afterwards is extrapolated to 60dB assuming that the part of the decay that we used is representative for the entire decay. It is common practise to specify the decay used (T20, T30, etc.).
One way of checking the consistency is by comparing e.g. T20 and T40. Any discrepancies between the two will normally originate from a non-linear decay (when plotted as a graph with a logarithmic level scale). All the Norsonic building acoustics instrumentation currently available has this capability of presenting at least two ways of calulating the decay simultaneously.



Case of the measurement time for T60