MULTISENSOR INPUT FOR A SPEECH ENHANCEMENT EQUIPMENT: 
SYSTEM ARCHITECTURE AND EXPERIMENTAL RESULTS (+)

Livio BUSNELLI, Neviano DAL DEGAN and Luigi VETRANO <**>

ABSTRACT

Speech enhancement techniques based on two inputs algorithms are completely inadequate if ambient noise reveals strong incoherence. On the other hand, enhancement obtained by using spectral subtraction methods presents several inconveniences while also being rather complex. The herein presented equipment and the relevant algorithms use an array of input sensors to pick-up the speech in a noisy environment; appropriate signal processing permits to exploit both the spatial coherence of the speech and incoherence of noise. The signal to noise ratio of the output signal shows a very good accordance with that of the theoretical model. The processing load can be undertaken by currently available DSP technology.

1. INTRODUCTION

A recent work [1] carried out in the framework of mobile radio telephony research area demonstrated that a two components model is quite adequate to describe the short term properties of the acoustical noise onboard a moving car. A strongly localized source, the running engine, shows a very high spatial coherence while the effect of wind, tires and other undefined factors results in a broadband incoherent noise. The literature [2] presents very good solutions for the task of cancellation of the first component; its coherence in fact suggests the use of two microphones one of which, located faraway from the speaker's mouth, can capture a noise which, after proper filtering, can be subtracted from the signal picked-up by the other microphone so providing for an improvement of the signal to noise ratio. This technique fails completely in presence of incoherent noise; other methods based on spectral estimations of the noise during non-voice intervals and subsequent subtraction [3] show several inconveniences and also a high complexity.

The technique presented in this paper takes advantage of both the spatial coherence of the voice and the incoherence of the noise by the adoption of a suitable array of N microphones properly positioned in front of the speaker [4].

2. THE MODEL AND RELEVANT THEORY

Figure 1 shows a possible schematization of the process of picking-up the speech, uttered in a noisy environment, by means of an array of sensors. It is assumed that the speech is a localized source; this means that signals s1, s2, ... , sN are different but perfectly coherent. It is also assumed that speech does not suffer reflections by surrounding walls (or realistically the reflected power is negligible).

Finally, noise is assumed to be incoherent, that is the coherence function between every couple ni, nj lies near zero on the useful bandwidth. Leaving to the reference [4] the detailed description of both the mathematics and the proposed algorithm let's here summarize the results:

- the summation process adds-up the amplitude of the useful signals s1, s2,..., sN whilst the power is added only for noise signals n1, n2,..., nN; the result is an improvement of the signal to noise ratio of \(10\cdot \log(N)\) dB;

- to obtain that gain, the useful signals must be added exactly in phase, that is a compensation for different path delays must be inserted before summation, and noise must be uncorrelated which means that sufficient distance must be guaranteed among microphones [1];

- the sampling frequency of the system must be high enough (64 KHz at least) to ensure a quasi-unity transfer function on the useful bandwidth (4 KHz);

<*> ITALTEL SIT - Central Research Laboratories (+) Patent pending 
-20019 Settimo Milanese - Milano (Italy)
the differential delay estimation process must be robust (against the noise) and adaptive in order to follow the movements of the speaker's head (cross-correlations over a suitable range of lags are computed among the reference signal and the other input signals every 200 ms).

While a complete dimensioning of the system has to be tailored on the specific environment, taking obviously into account the necessary trade-offs, a reasonable set of system parameters has been derived from simulation work and it is reported in Tab. I.

3. SYSTEM ARCHITECTURE

The architecture is based on a modular approach. A central module, containing the processing blocks and the time base generator, handles N peripherals modules each connected to its own sensor.

The block diagram of the system architecture is shown in Fig. 2 where:
- ICU (Input Channel Unit) is devoted to perform the analog signal conditioning, the A/D conversion followed by differential delay compensation;
- DELEST (Delay ESTimator) evaluates the differential delays between the reference signal and the other N-1 inputs;
- ADDFIL (ADDer and FILter) performs summation on the N signals coming from ICUs, then makes filtering, decimation and D/A conversion;
- ADGEN (Address GENerator) provides for the synchronous write addresses of the N delay lines.

Two major buses are employed in the system: the DEBUS for transportation of data among ICUs and DELEST, and AFBUS which carries data from ICUs to ADDFIL. In addition to the aforementioned buses there is the address bus WRA supplied to ICUs by the centralized unit ADGEN.

Two signal processors (TMS320 of Texas Instruments) constitute the core of DELEST and ADDFIL units respectively.

Looking at Fig. 3, the basic algorithm may be described as follows. The first part T1 of the interval Tu (see Fig. 3 a) is dedicated to the acquisition process: at the sampling rate 1/Ts DELEST moves a sample from each input channel to its data memory. The length T1 of the acquisition process is suitable for a correct delays estimation. At the end of T1 DELEST evaluates the differential delays by computing the N-1 cross-correlation functions, process which takes (N-1) Tcc.

The block ADDFIL works with a frame length Td (see Fig. 3 b) which depends on the used decimation factor. In the first part Ta of such interval the input process is carried out by adding the N "in-phase" samples coming from ICUs. In the next interval Tf, ADDFIL makes filtering and decimation.

The remaining Tg1 and Tg2 constitute guard intervals.

Input Channel Unit. The Input Channel Unit (ICU), as shown in Fig. 4, consists of a low pass antialiasing filter, an amplifier, a 12-bit A/D converter, a 256-word x 12-bit RAM acting as delay line (DELIN) and a set of latches to connect these components to the buses.

The sampling rate of the A/D converter is 64 KHz. This choice also allows for the use of a simple low pass filter. The amplifier is set to adjust the signal dynamics to the A/D converter input specifications.

Unit's operating principle is described in the following. The conversion is started by the signal SOC while simultaneously the DELIN write (WRA) and read (RDADDR) addresses are incremented. When conversion is completed, the sample is transferred into latch DEGATE and is also written into DELIN by enabling WRGATE. The delayed sample is read from DELIN by enabling RDGATE and then transferred into AFGATE. When DELEST wants to read a new sample of the analysis window, it addresses DEGATE as an input port by means of signal E1 and then transfers the sample into its memory via DEBUS.

ICU is also connected to ADDFIL through AFBUS using AFGATE which is enabled by ECI signal supplied by ADDFIL. The CNTRL block generates the control timing signals necessary for proper operation of the unit.

Adder and Filter Unit. The ADDer and FILter unit (ADDFIL) is shown in Fig. 5. It consists of the following parts: a 128-word x 12-bit dual port RAM (LUNGMEM), an address generator (ADGEN), a DSP unit, a decoder to address external ports, a 12-bit D/A converter, a low pass filter.
This block implements the summation and the decimation processes. Using ADRGEN, the N delay compensated samples coming from ICUs, are stored in LUNGMEM using N consecutive locations. Write operation is performed at N times the sampling rate. LUNGMEM is used by the DSP as its external data memory and is accessed through the local data bus. In such way the processor can read the N "in-phase" samples, add them and store the result in the on-chip data RAM (block B1). The on-chip data RAM block B0 is used to store the decimation filter coefficients. Decimation begins as soon as the processor has arranged a suitable number of data so that it can produce one output sample. The decimation output is sent to latch OUGATE that is handled as an output port by the DSP. OUGATE supplies a direct digital output and, by feeding the D/A converter, an analog output too.

Delay Estimator Unit. The DELay ESTimator unit (DELEST) consists of a DSP with its own program and data memories and two decoders to address external I/O ports. The DELEST signal processor communicates with the various ICUs via the DEBUS, the enable input signals (E1) and the enable output signals (EO1). DELEST works in three main states: acquisition mode, computation mode and monitoring mode. The transitions between these states are made using three interrupt levels: sample interrupt (Ts), frame interrupt (Tu) and monitoring interrupt (asynchronous). At the frame interrupt the DSP goes in acquisition mode by addressing the N ICUs to take the incoming samples and store them in its RAM. Then the DSP goes in computation mode where it loads the reference and delayed signals in the internal RAM (block B0 and B1 respectively), computes the cross-correlation functions over a suitable number of lags [4] and by means of them evaluates the differential delays. These estimations are the base for a proper setting up of the RDADDR signal (EO1). DELEST

4. COMPLEXITY EVALUATION

An evaluation on the processing power and memory requirements for the two main modules is given hereafter:

- DELEST unit. As shown in [4] the function chosen for the delay estimation is the cross-correlation, that is:

\[ C(k) = \sum [x_i(p) \cdot x_j(p-k)] \quad \text{for} \quad p = 1, L ; k = -M, M \]

Considering M=96, L=1024, and 1/Ts=64 KHz, the required processing power is about 200 K operations/channel and the external memory need is about 1.3 K words/channel.

- ADDFIL unit. Assuming a decimation factor of 8 and N input channels the required processing power, Pp, per output sample is:

\[ Pp = 8 \cdot N \cdot Kf \]

where Kf is a constant value depending on the filter length. In our case three stage FIR filtering process is employed with respectively 7 taps working at 64 KHz, 11 taps working at 32 KHz and 64 taps working at 16 KHz so that a figure of Kf=300 operations per output sample is achieved. This process only requires an external buffer of 8•N words (12 bit) which corresponds to the block LUNGMEM described before.

- ICU constraints. Using one TMS32020 DSP for DELEST and another for ADDFIL the former can handle only 5 input channels while the latter can handle 16 or more channels. Using a TMS320C25 for DELEST unit the number of input channels handling capability remains 9 ICUs. Because of this unbalance, the development of a more powerful algorithm for the delay estimation should be studied to improve the operating speed of DELEST so that the number of 16 input channels can be reached without modifications on hardware structure.

REFERENCES

Fig. 1. Model of the input system

Fig. 2. System architecture

Fig. 3. DELEST's (a) and ADDFIL's (b) time diagram

Fig. 4. Input Channel Unit block diagram

Fig. 5. Adder and Filter Unit block diagram

Tab. I. System parameters