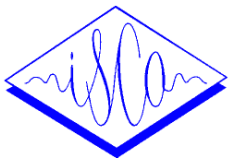

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

VOCAL PAUSES IN TEACHING: STATISTICAL ANALYSIS AND APPLICATIONS

E. Angeleri†, M. Barsotti†, L. Mazzei†, L. Vetrano‡, R. Volpentesta†
†Dipartimento Scienze della Informazione, Università di Milano,Via M. da Brescia 9, 20133 Milano, Italy
‡ITALTEL–SIT laboratorio DSP, Cascina Castelletto, 20019 Settimo M.se MI, Italy

ABSTRACT

The distribution of pauses in universitary lesson is presented. The aim of this study is the definition and specification of a system for voice–data interpolation on switched telephone networks. The results obtained shove to the realization of a transmission protocol and a hardware architecture achieving the target. This paper gives an overview of the studied approach; some suggestions for different applications and further improvements are also provided.

INTRODUCTION

The distribution of voice pauses in teaching activity has been exploited for the realization of multimedia teaching systems. In this paper the possibility of using a switched voice channel for data and voice transmission is investigated. This system is suitable for use in a teaching environment where only teacher's voice and simple messages for remote data generation on a computer must be sent. With adequate equipment it is possible to handle steady images also.

The pauses can be used for data transmission [4], [7] or for voice compression [6].

This paper is divided in four sections according to the logical steps undertaken in the experimental phases.

In section one the pauses percentage and its characteristics in an universitary lesson are estimated. In section two are studied the effects of pauses elimination and talkspurts packing in order to increase the compression rate of some voice coders; the results are exposed for some of the most common.

A protocol for voice and data interpolation is finally given with a hardware architecture specifically conceived for this need. It is worth noting that the last solution could be useful also in other applications.

I THE PAUSES IN TEACHING ACTIVITY

A lot of literature during the last 30 years deals with the use of pauses during telephone conversations ([5], [2]) with interesting statistics on this subject. Unfortunately, to the best of our knowledge, nothing similar seems to have been done for teaching activity. In order to investigate this problem some recordings of universitary lessons are studied.

In these experiments some teachers used a blackboard, others a overhead projector for trasparencies holding a recorder with a tie microphone.

Some parts of these registrations were randomly chosen for a whole duration of 20 minutes and converted in digital 16 bit samples with 8 KHz sampling rate at the ITALTEL–SIT DSP laboratory. These parts are used as input to a program which implements an algorithm for voice–silence discrimination. The algorithm works on a 10 ms basis frame giving as output a Voice Activity Detection (VAD) flag.

The De Souza autoadaptative algorithm has been used in the experiments for discriminating between voice and background noise [3]. The procedure is based on a pattern recognition approach and avoids the problem of subjective prefixed thresholds tuning commonly used in literature [1].

After a training period the frames in the De Souza algorithm are classified in speech or silence by means of five parameters: energy in Db, Zero Crossing Rate function, autocorrelation coefficient, number of points where there is signal change of direction and a signal measure differenziated and normalized [3]. Due to the statistical nature of the measure a confidence level is needed.

The training procedure with the precision parameter proposed in [3] hinders in some situations the use of statistical test because of classification errors. In the case considered here the problem has been avoided by increasing the confidence level (i.e. 0.03 or 0.05). This gives satisfactory results, but for longer recordings this technique could possibly give the same problem.

The algorithm output is a sequence of zero and ones representing the silence or voice frames. Possible isolated spurts of 10 ms, caused by the background noise, must be eliminated from this sequence because they could trick the VAD.

The following parameters are excerpted from these sequences: pauses percentage (its complement is the voice activity percentage), number of pauses per minute, talkspurts and pauses average durations in seconds.

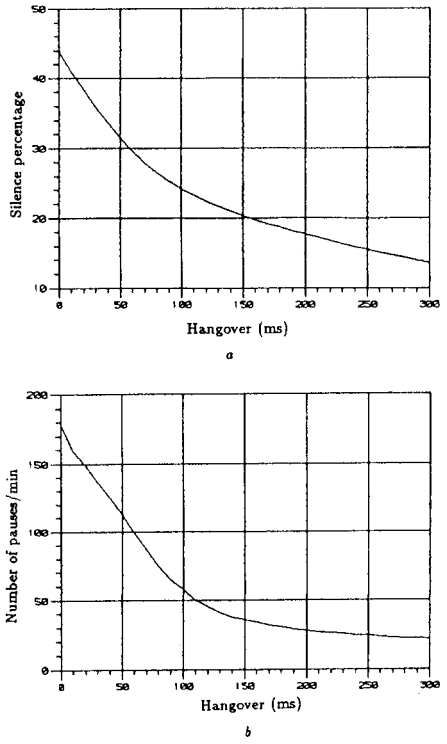

Fig. 1: silence percentage (a) and number of pauses by minute (b) as function of Hangover.

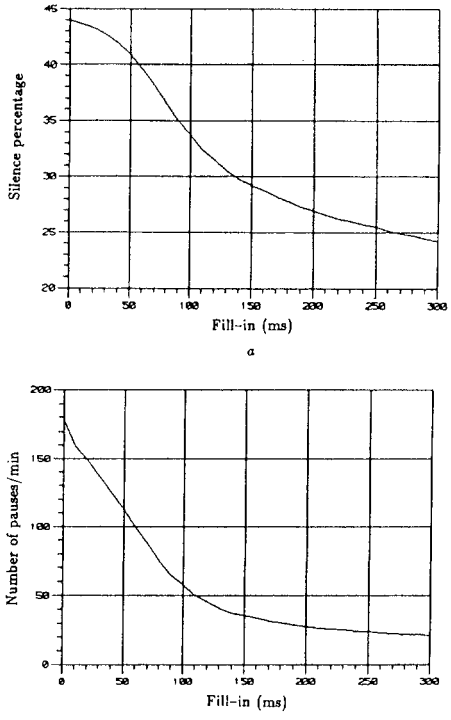

Fig. 2: silence percentage (a) and number of pauses by minute (b) as function of Fill-in.

All values are measured as a function of hangover and fill–in time in a range from 0 to 300 ms. Figures 1 and 2 contain a selection of more interesting results.

Some considerations can be drawn from these results. The increase of hangover and fill–in time implies a decrease of silence percentage and number of pauses, with a rapid evolution up to a value of 150 ms when all intersillabic pauses are filled. Subsequently the decrease is less marked, due to the lack of short pauses; it becomes rapid again as soon as the hangover is so high as to fill the pauses between the phrases.

In table 1 the average values measured with and without hangover and fill–in are reported.

Table 1: Statistics of some experimental results.

	Zero Hangover	Hangover 200 ms	Fill–in 200 ms
Pauses	43.97%	17.68%	26.91%
Pauses/minute	178.51	27.83	27.68
Talkspurt dur.	0.19 s	1.85 s	1.66 s
Pauses dur.	0.15 s	0.37 s	0.57 s

The values are highly subject dependent; to understand the reasons of these differences a study about pauses and talkspurts density has been performed. This study has ascertained that short pauses (less than 200 ms) are common to all subjects while longer pauses (greather than 400/500 ms) differs from subject to subject. The talkspurts density instead are not so variable (figure 3).

From these results short pauses and talkspurts durations could be dependent on the structure of the particular language used. This aspect is still under investigation.

II OPTIMIZATION OF VOICE COMPRESSION METHODS

This section shows the application of the VAD algorithm to a voice coder in order to improve the effective compression rate.

The main idea is to modify the signal duration by removing the pauses and then coding this time–compressed signal with a traditional codec. On the receive side, after decoding, pauses are reconstructed according to information stored in the transmitted pattern.

This goal has been achieved via a simulation where the input voice frames were digitalized as previously described creating two output files. One with the voice without pauses (*compressed file*) generated by the De Souza algorithm and another with VAD flags describing the presence or absence of voice on the 10 ms frame basis (*file pattern 0–1*). Every bit represents 80 samples corresponding to 10 ms of input signal (the sampling rate is 8 KHz).

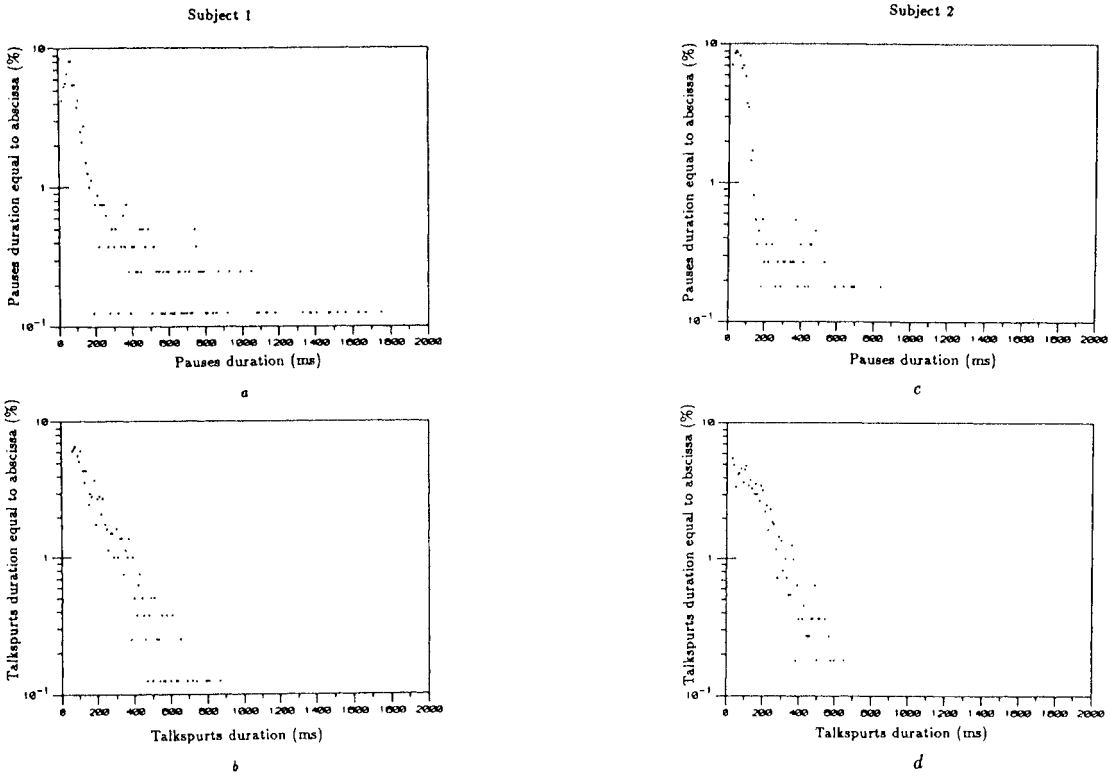

Fig. 3: pauses duration density (a and c) and talkspurts duration density (b and d) for two subjects.

The *compressed file* is first processed by a voice coder and then decoded; finally, pauses are reconstructed by means of *file pattern 0–1* formely created. The whole system is shown in figure 4.

In early stages of the experiments the throaway, hangover and fill–in time was set to zero and no background noise was produced during the pauses. The effect was a poor voice quality, as expected, but the words were still meaningful. The annoying effect due to the presence/absence of the background noise is reduced by the introduction of artificial noise (*comfort noise*).

The same experiments repeated with hangover of 50 ms and a simulated gaussian white noise give better results.

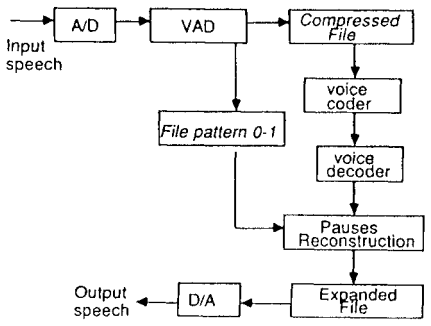

Fig. 4: scheme adopted for testing the optimization of voice compression methods.

We tested the algorithm in combination with a RPE–LTP speech coder, a CELP coder and a LPC–10 coder, according to the scheme of figure 4.

In all cases the quality of the whole process is satisfactory.

The presence of a slight reverberation effect on the reconstructed speech is typical of such kind of processing and some solutions are still under investigation.

III A TRANSMISSION PROTOCOL FOR VOICE–DATA INTERPOLATION

In the previous sections the possibility of voice–data interpolation has been established. In the present one a protocol achieving this goal is proposed.

In figure 5 the packet structure for voice and data interpolation is presented. The packet length is fixed and, as shown in the figure, can be partitioned into two main parts. The first (A) contains the codification of compressed voice during a temporal interval equals to the total packet duration.

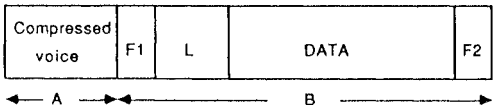

Fig. 5: packet structure proposed for voice–data transmission protocol.

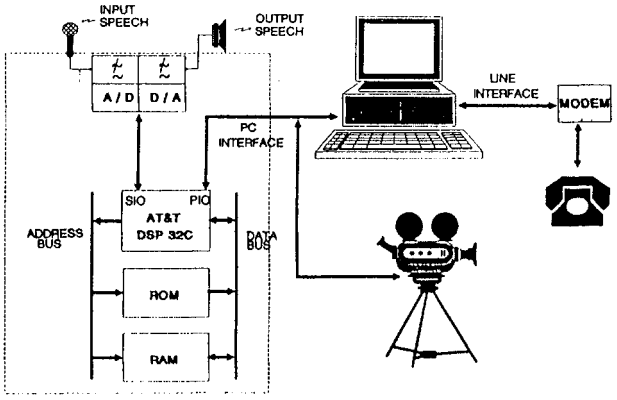

Fig.6: hardware architecture for voice–data interpolation

In part B two flags (*F1, F2*) delimit the beginning and the end of non voice segment of the packet (the voice part has a variable duration).

The information for the reconstruction of the pauses is contained in section *L* of the packet, where the bit sequence is interpreted as the presence or absence of noise during the period corresponding to the voice duration. For istance, if the voice utterance lasts 80 ms and every voice frame is 10 ms then the sequence 00001110 means that a pause of 40 ms is followed by a 30 ms talkspurt and a 10 ms pause (notice that the atomic unit is the frame length).

The voice transmission in this scheme implies a fixed delay, dependent from the packet dimension, which should be wisely chosen. In order to verify the effective amount of space available for data transmission, a pattern of voice derived from the recorded lessons has been studied. The pause percentage with a 50 ms hangover turns out to be of about 30%.

IV A HARDWARE ARCHITECTURE FOR VOICE–DATA INTERPOLATION

The world telecommunication network is in the process of being digitized for both improving the quality and introducing new services. In the meantime, each new service must be checked on the analog telephone network in order to reach potentially everyone.

Figure 6 shows the approach followed for the experimental phase of the voice–data interpolation system. The processing board, based on the DSP32C (AT&T), manages speech samples coming from the A/D converter and data stored into the PC memory.

The DSP32C is a 32–bit floating point DSP running at the speed of 50 MHz; its computational power is 12.5 MFLOPS with a peak of 25 MFLOPS. Using the DMA (direct memory access) capability of the DSP, we have that while the speech frame "K" is coming into the memory via *SIO* (Serial I/O) interface, the frame "K–1" is being elaborated by the DSP and the frame "K–2", already elaborated, is coming out via *PIO* (Parallel I/O) port.

The programmer's only concern is to ensure that the whole program (speech coding, voice activity detection, protocol handling and data packing) completes the execution in time to process the next block of speech. The *PIO* port is also used by the DSP for getting data from the PC.

According to the VAD output, the DSP decides whether to transmit to the other end elaborated speech or compressed data.

The same architecture is also suitable to check the time scale compression of speech samples in order to join the pauses together. This strategy allows a more efficient use of the pauses, as shown before, while preserving the speech quality.

Our architecture does not allow direct input of images from a camera into the PC memory, a commercial *Matrox* board connected through the PC bus has been used. Images are compressed off–line by the PC microprocessor and then transmitted according to the VAD decisions.

CONCLUSIONS

This paper has shown that there are enough pauses in an universitary lesson to justify their use for data transmission. A further improvement could be achieved by means of pauses suppression followed by voice coding. A simple protocol and a hardware architecture were finally presented which achieve this goal.

It should be noted that the longest pauses could also be used for segmenting the lesson into logical units; in fact other experiments showed that in most cases a long pause ends a meaningful sentence.

REFERENCES

[1]: **P. Brady**, *A Technique for Investigating On–Off Patterns of Speech*, BSTJ, , Jan 65

[2]: **P. Brady**, *A Statistical Analysis of On–Off Patterns in 16 conversations*, BSTJ, , Jan 68

[3]: **P. De Souza**, *A Statistical Approach to the Design of an Adaptive Self–Normalizing Silence Detector*, IEEE ASSP, vol. 31, Jun 83

[4]: **M.J. Fischer**, *Data Performance in a System where Data Packets are Transmitted During Voice Silent Periods. Single Channel Case*, IEEE COM, vol. 27, Set 79

[5]: **J.M. Fraser, D.B. Bullock, N.G. Long**, *Over–All Characteristics of a TASI system*, BSTJ, , Jul 62

[6]: **K. Gan, R.W. Donaldson**, *Adaptive Silence Detection for Speech Storage and Voice Mail Applications*, IEEE ASSP, vol. 36, Jun 88

[7]: **E. Lyghounics, I. Poretti, G. Monti**, *Speech Interpolation in Digital Transmission Systems*, IEEE COM, vol. 22, Set 74