Audio Processing Solution for Video Conference Based Aerobics

Magnus Berggren[a], Louise Stjernberg[b], Fredric Lindström[c], Ingvar Claesson[a]

[a]Blekinge Institute of Technology, Department of Signal Processing, SE-37225, Ronneby, Sweden
[b]Blekinge Institute of Technology, School of Health Science, SE-37179, Karlskrona, Sweden
[c]Limes Technology AB, Box 268, SE-90106, Umeå, Sweden

***Abstract*—In this paper an audio processing solution for video conference based aerobics is presented. The proposed solution leaves the workout music unaltered by separating it from the speech and processing each signal separately. The speech signal processing is also performed at a lower sample rate, which saves computational power. Real time evaluation of the system shows that high quality music as well as a good two-way communication is maintained during the aerobic session.**

I. Introduction

Video conference equipments are well known CE-products which in recent time have become more and more used as a way to communicate over long distance. Today's increasingly available bandwidth and data transfer speed allows for good video and audio quality.

Audio solutions for video conference are in general targeted for two-way speech communication. As in any hands-free communication application the loudspeaker signal is inevitably picked up by the microphone, generating an acoustic echo. Echo cancellation or echo suppression is typically used to remove the acoustic echo [1][2].

The main difference between an aerobic application and a conventional video conference application is that for the aerobic case continuous music is present. Further, high listening comfort requires a wideband music signal and that no damping is applied to the music signal.

II. System

The setup used for the video conference based aerobics is illustrated in figure 1. The participant side consists of a set of speakers, an omni-directional microphone, a box with the application specific audio processing and a video conference system with codec and internet connection. The instructor side consists of speakers, a wireless headset microphone commonly used by aerobic instructors, a music device (e.g MP3-player), a box with the application-specific audio processing and a video conference system. On the instructor side the music is added to the speech signal after echo cancelling and this signal is sent to the loudspeakers and to the participant side.

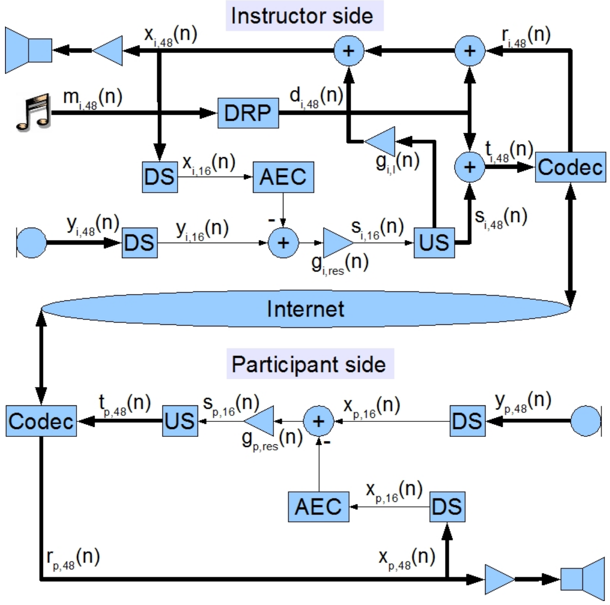

Fig. 1. Schematic view of the audio signal paths, thick lines denote 48 kHz signals and thin lines denote 16 kHz signals.

III. Mixed frequency

Computational power is a limited resource in CE-products, implying a demand for low complexity applications. In the required audio processing the most computationally heavy task is the echo cancellation. By using different sampling frequencies for the music and the speech, 48 kHz and 16 kHz respectively, high quality music and reduced speech echo cancellation complexity is achieved. This processing is illustrated in figure 1. The instructor side loudspeaker signal $x_{i,48}(n)$ will be

$$x_{i,48}(n) = r_{i,48}(n) + g_{i,l}(n)s_{i,48}(n) + d_{i,48}(n), \qquad (1)$$

where n is the sample index, the subscript 48 indicates the sampling frequency 48 kHz and the subscript i is for instructor side. $r_{i,48}(n)$ is the received signal from the codec, $s_{i,48}(n)$ is the processed speech signal and $g_{i,l}(n)$ is a damping to eliminate howling, see section IV. The signal

$$d_{i,48}(n) = L\left(m_{i,48}(n)\right) \qquad (2)$$

is the dynamic range processed (DRP) music signal, where $L()$ is the function for dynamic range processing, see section V. By

Thanks to the Swedish Knowledge Foundation (KKS) for funding.

downsampling (DS) the microphone and speaker signal to a lower sampling rate, acoustic echo cancellation (AEC) with reduced complexity can be performed. The echo cancelled near-end speech is given by

$$s_{i,16}(n) = g_{i,\text{res}}(n)\Big(y_{i,16}(n) - \mathbf{x}_{i,16}^T(n)\hat{\mathbf{h}}_{i,16}(n)\Big), \quad (3)$$

where $y_{i,16}(n)$ is the 16 kHz microphone signal, $x_{i,16}(n)$ is the 16 kHz loudspeaker signal used in the vector $\mathbf{x}_{i,16}(n) = [x_{i,16}(n), x_{i,16}(n-1), \ldots, x_{i,16}(n-P+1)]^T$, $\hat{\mathbf{h}}_{i,16}(n) = [\hat{h}_{0,i,16}(n), \hat{h}_{1,i,16}(n), \ldots, \hat{h}_{P-1,i,16}(n)]^T$ is the estimated room impulse response of length P and $g_{i,\text{res}}(n)$ is the residual echo suppression gain used to reduce the echo even further. The residual echo suppression gain is calculated in the same manner as the fullband residual echo suppression described in [3]. The near-end speech is then upsampled (US) to 48 kHz and the signal transmitted to the codec becomes

$$t_{i,48}(n) = s_{i,48}(n) + d_{i,48}(n). \quad (4)$$

On the participant side, p, the processing is carried out analogously.

IV. Instructor Voice Monitoring

The voice of the instructor is directly fed to the loudspeaker on the instructor side to give voice monitoring to the instructor. This will make a loop-gain larger than one possible, which could result in howling, especially when the instructor is close to the loudspeaker. The gain should therefore be adjusted with a damping to ensure loop stability. The feedback gain, $g_{i,l1}(n)$, can be calculated as

$$g_{i,l1}(n) = g_{\text{aec}}(n)\frac{\overline{y}_{i,16}(n)}{\overline{x}_{i,16}(n)}, \quad (5)$$

where $\overline{y}_{i,16}(n)$ is a time average of the microphone signal, $\overline{x}_{i,16}(n)$ is a time average of the loudspeaker signal and $g_{\text{aec}}(n)$ is a gain that corresponds to the AEC performance. Another way of estimating the feedback gain is to calculate it from the impulse response as

$$g_{i,l2}(n) = g_{\text{aec}}(n)\sqrt{\sum_{k=0}^{P-1}\left|\hat{h}_{k,i,16}(n)\right|^2}. \quad (6)$$

The final loop damping uses a combination of the estimates in eq. 5 and eq. 6 according to

$$g_{i,l}(n) = \frac{1}{\max\left(g_{i,l1}(n), g_{i,l2}(n)\right)}. \quad (7)$$

V. Dynamic Range Processing

Practical experience shows that different music devices have different maximum output volume. To adjust for this the signal has to be amplified so that the music device with the weakest output gives a sufficiently strong signal. However the amplification can cause music devises with strong output volume to be digitally clipped due to saturation. The music will then sound distorted. To avoid this a limiter is necessary. Due to the fact that music consists of slow beats and parts where the music is louder or weaker, the limiter has to be very slow so that no apparent adjustment of the gain is noticed.

Fig. 2. The participants in the middle of an aerobic session.

VI. Real-Time Evaluation

During a test period of 4 weeks, the audio processing application was tested by a group of office workers. The aerobics was targeted for work-dressed participants in a 10-15 minutes session. The participants as well as the instructor were able to communicate with each other the whole session and high music quality was maintained.

Figure 2 shows an aerobic session in full action.

VII. Conclusion

A conventional video conference application was modified to function in a video conference based aerobics situation where continuous music, in addition to speech, is present. The difference in music player outputs and instructor voice monitoring called for extra processing functionality. In addition, the proposed audio processing solution reduces complexity by processing the speech and music signals separately and at different sample rates. During real time evaluation of the system it was shown that a good two-way communication could be maintained during the aerobic sessions.

References

[1] S. Haykin, *Adaptive Filter Theory*, 4th ed., Prentice, 2002.

[2] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, Wiley, 2004.

[3] C. Schuldt, F. Lindstrom and I. Claesson, ”Combined Implementation of Echo Suppression, Noise Reduction and Comfort Noise in a Speaker Phone Application”, *Proceedings of International Conference on Consumer Electronics*, Las Vegas, NV, 10-14 Jan. 2007