Audio Quality and Acoustic Echo Issues for VOIP on Portable Devices

Giorgio Zoia, Alain Sturzenegger
EyeP Media SA
Yverdon-les-Bains, Switzerland
{giorgio.zoia, alain.sturzenegger}@eyepmedia.com

Olivier Hochreutiner
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
olivier.hochreutiner@epfl.ch

***Abstract*—The control of the overall quality of service for VOIP applications depends for sure on features and capabilities of the network layer but at the same time, and especially in the case of low-resource, portable devices, it relies on smart flexible signal processing and control tools that allow tuning the usage of computational resources to obtain the best possible quality for a given system. Audio signal I/O (acquisition and rendering), packet loss concealment and acoustic echo detection and suppression play a fundamental role, since good algorithms are rather demanding for calculations but their presence is often necessary to compensate low-cost, average-quality platforms and interfaces. Some measures and remarks on these platforms and interfaces are presented in this paper, together with a set of implemented solutions providing a remarkable improvement in the mean opinion score of a typical VOIP conversation for different families of PDAs and Smartphones.**

Keywords-VOIP; audio quality; voice quality; acoustic echo; signal processing;

I. INTRODUCTION

The fast spreading of Voice-Over-Internet-Protocol (VOIP) telephony and the increasing capabilities of low-power devices make more and more feasible and attractive the scenario of voice (and video) communications based on Wi-Fi connections for portable devices. In order to obtain a satisfactory quality of service, a number of challenging issues arise concerning audio interfaces and sampling rate, Packet Loss Concealment (PLC), and especially acoustic echo control. Acoustic echo appears because the signal emitted by the loudspeaker is captured by the microphone and then returned back: as a consequence, the remote speaker hears itself again.

Audio quality and audio enhancement issues are particularly important in hands-free configurations, where one or two small external loudspeakers are used at a volume that can make voice audible within a certain distance. We are particularly interested in this configuration for many reasons, among which: a) it represents an increasing center of interest especially when video telephony (V^2OIP) is considered; b) it is requested for calls between small groups or for conferencing; c) it may be used in cars; d) some PDAs do not provide a different audio device for sound output. For all these reasons the processing tools are requested to operate in such condition. In addition, it is desirable that signal processing algorithms are as far as possible versatile, in the sense that every algorithm can be tuned to work at best on a particular device, but its core must be generic enough to allow an extremely fast porting and basic performance in a wide range of conditions. This is a requirement in clear contrast with optimization of resources.

In portable devices resources are always limited; for this reason and for the restrictions in the numeric format, signal processing on such devices is often a challenge. In this paper we present an approach to audio quality enhancement and echo control specifically targeting VOIP applications on portable devices. Results are positive and already represent, to our knowledge, a step ahead of a majority of currently available tools in the domain. The presented signal processing toolset includes good quality sampling rate conversion in combination with PLC technology and echo suppression; the flexibility of the algorithms and of the softphone tool where they are integrated allows a good adaptation on all the different devices where it is being tested, in terms of trade-off among required resources (for the device and the network), short term audio quality (both local and remote), and long term perceived conversation quality; the control of all these features is necessary to provide the best possible overall quality of service given a certain set of conditions and setup.

II. AUDIO SIGNAL I/O AND ERROR CONCEALMENT

A. Audio Signal I/O on Portable Devices

Controlling the overall audio quality is often a rather challenging problem in portable devices. In fact, many of them are rather undocumented machines characterized by an objective difficulty in accessing fundamental parameters such as microphone sensitivity, speaker selection, and sampling rate conversion. Furthermore, their integrated audio devices are quite often low in quality, characterized by highly non linear behaviors and poor dynamics.

Fig. 1 shows measurements for two typical cases in current commercial PDA devices (very similar results can be obtained with many other devices, like e.g. Qtek 9100 and Qtek 8310, but they have been omitted to improve readability). These measurements have been obtained feeding white noise to the speakers, carefully avoiding saturation in loudspeakers (volume 3/5), and recording the echo signal. Spectral estimation techniques have been used to calculate the curves. Therefore the plotting represents the overall echo-path response that can be expected for those devices.

1-4244-1039-8/07/$25.00 ©2007 IEEE

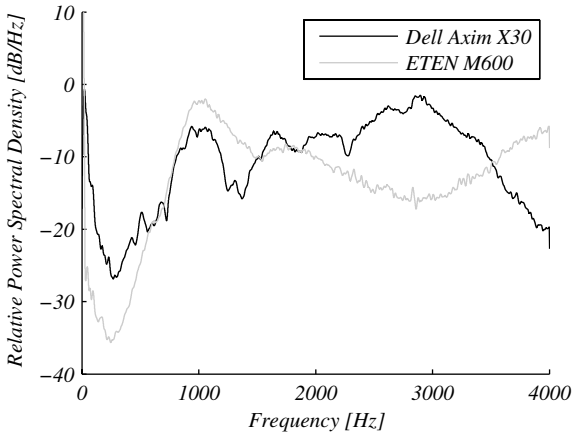

Figure 1. Typical frequency responses of echo paths in portable devices. The curves cumulate the effect of speaker and microphone in hands-free configuration.

At the same time this result provides a good idea of the typical distortion that can be expected in a portable-to-portable communication for the path from the microphone at one end to the speaker at the other end.

The combination of these often poor I/O devices with low quality drivers is particularly disturbing for at least two different reasons.

First of all, the internal data stream in telephony is normally at a sampling rate of 8 kHz (narrowband) or 16 kHz (wideband) and since these frequencies are in most cases not natively supported by hardware samplers, a conversion automatically takes place in the device drivers which is usually bad in quality and precision because of the computing resource limitations. Signal distortions and aliasing noise are generally observed as a consequence of this poor signal processing when working, like it is in our case, with the widespread Windows CE operating system (in its different versions and flavors). But the situation is not so much different in other cases that have been considered and tested recently.

Secondly, the control of disturbance elements like acoustic echo requires a continuous analysis of the decoded signal fed into loudspeaker(s) and of the microphone signal, as far as possible maintaining a proper synchronization between them. Frequency shifts and signal distortion coming from output and input audio devices can be so important that many generic echo control algorithms are unable to provide useful results: this is particularly true for adaptive algorithms but also for energy-based cue trackers, like it will be presented later in section III.C.

A first fundamental stage included in our family of products consists of a set of methods to increase the quality of the audio signal capture and rendering, bypassing as far as possible the native OS SDK by leaving to it only the basic low level operations. The most important of these methods provides a good quality resampling at a precise rate; this improves the audio quality by reducing distortion and aliasing noise, while at the same time it allows a much more precise synchronization between output and input (when coupling from speaker to microphone is observed), avoiding frequency shifts and delay oscillations.

At the same time a number of cheap IIR filters has been designed aiming at the suppression of extremely distorted (and then more disturbing than useful) signal bands or at the compensation of losses in other bands due to highly non-linear frequency responses (see again Fig. 1).

B. Packet Loss Concealment

When VOIP tools are used with desktop or laptop computers, especially at home or office, they rely on cabled interned connections or on Wi-Fi connections that are optimally setup for signal reception. Usage of VOIP tools on mobile devices implies a wireless connection with, on the average situation, lower signal strength. The packet loss ratio on portable devices is consequently higher than for computers in most cases, and this does not help the perception of a good quality conversation.

Many algorithms exist for packet loss concealment, and some of them are standardized by telecom committees. ITU-T G.711 Appendix I [1] is probably the most used one, but its algorithm is specified in terms of floating-point energy-based pitch detection that is of course not usable for fixed-point architectures such as ARM cores. Other algorithms exist that can improve further quality but can be even more resource consuming [2]. Our solution consists instead of a G.711 Annex I porting to fixed-point arithmetic, further optimized with additional simplifications and approximations in the energy-based pitch detection for buffer re-synthesis. In spite of these approximations, experimental results have shown that the same pitch is calculated than the floating-point complete algorithm in 85% to 90% of cases; only 2-5% of estimations are wrong by more than 2 Hz, always in comparison to the floating-point algorithm.

III. ACOUSTIC ECHO DETECTION AND CONTROL

Acoustic echo appears when some of the signal emitted by the loudspeaker is captured by the microphone and returns back to the far end: as a consequence, the remote speaker hears itself again. Depending on the amount of echo and on the delay introduced by the network and by the software layer (application interfaces and drivers), this phenomenon can make the telephonic conversation unpleasant or, in the worst case, very difficult.

The party responsible for the echo does not suffer from the echo that is created at its side: the far-end speaker suffers instead from the echo created by the system at the near-end side. Completely hands-free telephony always creates an echo; that echo may be faint depending on the layout of the loudspeakers and microphone, but this is unfortunately not the case in practice.

Acoustic echo detection and control constitute a very complex and tricky domain of research. Work is ongoing on these topics since a few decades [3], but the continuous evolution in devices, platforms and use conditions always creates new challenges. In this paper we do not intend to present a comprehensive review of acoustic echo control, but

our scope is instead limited to those aspects and problems, with the corresponding solutions we found, that are inherent to the domain of portable devices.

A. Related Work

Acoustic echo can be detected and controlled in several different ways. The most widely used technology is probably the linear adaptive one, e.g. with an LPC filter adapted by NLMS (Normalized Least Mean Square). This technology works at best when conditions are stable; the best performance and robustness can be achieved working on block of samples in the frequency domain, being convergence already uncertain in case of very small adjustments of the reference signals (this is especially true for consumer electronic devices). Linear adaptive algorithms have the problem of (non-linear) echo residuals, which are often not acceptable if echo energy is approaching or even amplifying the speaker signal energy. A great majority of mobile phones integrates echo control algorithms of this type, as levels are usually sufficiently low and the device features are well-known and characterized with good stability.

Non-linear adaptive techniques exist, but their improvement over linear ones is often reduced to a very few dBs, at the price of much more relevant resource requirements (see e.g. [5]).

Finally, technologies based on a more or less intelligent gain control are often used, especially when echo levels are rather high, when available resources are limited or when the overall system presents challenging instabilities. In this family of algorithms, which can be labeled as gain switching algorithms, the typical problem is coming from voice chopping. Gain switching algorithms can be characterized by one or more different parameters; very promising approaches in this sense are those with perceptually motivated subband gain controls [6, 7].

Several different combinations of the above mentioned families of techniques are of course possible, according to the specific targets, desired quality, etc., like it is the case for the work presented in this paper.

B. Delay Estimation

In the following, we intend by "output time" the instant when sound samples are written to the device driver buffer and by microphone "input time" the instant when samples can be retrieved from the device driver buffer. Especially when being used in a hands-free configuration, but sometimes even with ear loudspeaker, a considerable feedback (echo path) exists in a portable device from the output spot to the microphone input due to different acoustic and mechanical couplings (see also next section). The delay between input time and output time introduced by this path varies considerably from one platform to the other according to different buffering strategies, OS device drivers and corresponding latencies; as mentioned above, this delay might even change in time due to drifts caused by the drivers themselves. It seems evident that a stage of delay estimation is fundamental to effectively deal with echo cancellation in a sufficiently general way, and that the acoustic path alone only accounts for a small portion of this overall delay (especially in portable devices the acoustic path is extremely small).

Delay may be estimated in many different ways, in time or frequency domain, often using energy or signal correlation techniques. Energy computation provides in many cases more reliable results; at the same time algorithms based on energy mean values, sometimes weighted on rather long delay lines, require a huge dynamic range, which is difficult to handle in fixed-point arithmetic with good precision. Instead, delay is initially estimated in our tool by a smart signal correlation technique based on two successive refinements, and this estimation is dynamically refreshed at a convenient rate around the initial value. To give an idea of the order of magnitude of the local portion of this delay, practical experiences made by careful measurements of the overall path (by mixing at the input the samples sent to the output and recorded later in form of echo) provided average times between 250 and 300 ms for widespread PDAs and Smartphones, running the Windows CE family of operating systems. A major part of this delay comes from long buffering needed to guarantee sound fluidity. Given the computational power of such devices, it is evident that any echo handling algorithm can not be implemented in practice without preliminarily estimating this delay, compensating it, and centering a smaller window around this estimated value for the successive stage.

Furthermore in mobile platform systems, even more than in fixed platforms, a buffering mechanism is necessary to compensate synchronization problems in data acquisition inside one thread when reading data from another one, like it can be the case between audio input (microphone-driven) and audio output (RTP-driven) data paths. Abrupt changes in the delay can be observed for instance in case of buffer re-alignment after a thread being long suspended; this can happen when e.g. the network adapter is blocked for some reason due to traffic or bad signal quality, or when the OS is not reacting fast enough due to CPU load. A continuous adjustment around more or less 1 buffer length is then required to maintain good delay estimation (i.e. ± 30 ms in our case).

C. Acoustic Echo Handling

Once the delay estimated, a scheme for echo control must be selected; some classic schemes can be found e.g. in [3] and [4]. In Fig. 2 an adaptive stage (Acoustic Echo Canceller, AEC) tries to converge by minimizing the error signal between the microphone input and its own output (when local activity is not detected). At the same time a suppression stage (Acoustic Echo Suppressor, AES), mainly based on a double-talk detection stage (DTD), is supposed to completely remove echo residuals in case of non optimal convergence of the adaptive stage.

In fact, in portable platforms it is rather difficult to obtain good results from an AEC stage alone. Since delay can be unstable, as said earlier, and devices always show a highly non-linear behavior, an adaptive filter in time domain (LPC filters or similar) often fails even in rather simple communication scenarios for this reason alone.

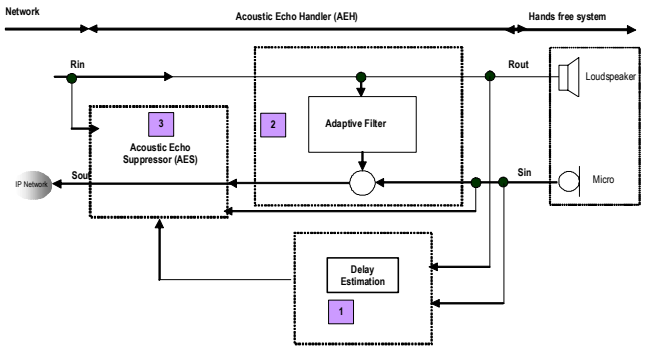

Figure 2. Implemented echo control module

Good adaptive solutions exist in frequency domain, like for instance multi-delay techniques presented in [4] and [8], but they present two main disadvantages: first, they are based on Fourier Transforms, which are known for being difficult to tame in fixed point; consequently, and secondly, computation in real numbers format is required, which is far from being affordable on ARM core processors.

It is wise to remember that in case of hands-free speaker configuration in portable devices, the echo coupling between speaker output and microphone input is extremely high and sometimes very difficult to control. To give some figures, measurements have been made with different volume settings for the power ratio between the microphone recorded signal and the speaker signal, with white noise as test input.

Results are presented in Table I for the same two devices of Fig. 1. Some couple of devices put in remote connection can generate a Larsen effect on background noise alone at volumes starting from 3-4/5. The small ear-speaker figure is reported in one case for comparison.

In such operating conditions, even a rather well working adaptive technology encounters serious problems, as it is easy to understand; the best results that could be obtained in practice by an NLMS-based technique and in particularly favorable situations did not provide more than 12 dB echo attenuation, far from enough especially in case of a single talker speaking.

To reach an acceptable quality of service even in such a very difficult environment, it must be accepted to start from the assumption that, in a normal conversation, speakers do not talk at the same time during the majority of the call duration. In this situation, just suppressing the echo by mean of an AES may be enough: this stage is consequently leading the echo control, by imposing extremely strong gain attenuations (60 dBs and more when the confidence is high) and relying on the adaptive stage only as an auxiliary support.

TABLE I. MICROPHONE/SPEAKER POWER RATIO

Speaker	AXIM X30	M600
Loudspeaker 3/5	-7.7 dB	-9.8 dB
Loudspeaker 5/5	+8.5 dB	-3.3 dB
Small earspeaker	n. a.	-33 dB

In the short excerpts of conversation when two sentences slightly overlap or both speakers talk at the same time the AES must let the signal pass to ensure the full duplex operation; in this case either the adaptive stage can continue to be operative to provide the best possible attenuation, or instead an intelligent, perhaps perceptually weighted, gain adaptation can be implemented to help reducing the echo perception, partially masked by a contemporary talk at the near end.

D. Proposed Implementation

If a DTD stage is an important element of an adaptive technique, allowing freezing of the coefficient adaptation when local activity is detected, it constitutes the very heart of an AES technique. In an extremely simplified model, DTD information is used as an input to a gain controller, which actually provides the echo suppression functionality.

The Geigel DTD algorithm [9] is probably the simplest, and then the lighter in terms of required resources, among existing effective solutions. At the same time, being simply based on a predefined threshold value, it reveals some limitations when the level of echo feedback is changing, like in case of speaker volume change. But indeed, by simply adding a Geigel stage, just through accurate tuning and allowing some constraints on the volume settings it is possible to transform a conversation quality from “bad” to nearly “fair”.

Better results can be obtained by slightly more sophisticated solutions. In our tool, advantage is taken of the correlation computed in the delay estimation part. The cross-correlation coefficient provides a measurement of the similarity between the loudspeaker and microphone signals, which tells if there is a situation of echo (high value), local speech (low value), or doubletalk (medium value). A normalization of the correlation coefficient is performed to make it independent of the signal energies.

An AES based only on cross-correlation gives acceptable results but it reveals some weaknesses during local speech activity, in particular at the beginning of sentences. To improve this situation with a further criterion, an energy-based local-speech detector has been added, which is more reactive.

In both cases, Geigel and cross-correlation, an additional protection mechanism for low energy syllables at the end of words and sentences has been added. This mechanism has a different behavior in presence of sections of different length to avoid problems with false alerts due to instant noise events.

Gain control is implemented by (pseudo-) intelligent agents providing a rather tunable decision module. Different decision criteria can be considered, as said above, with some of them impacting the quality of the final result more than others. More or less rules and more or less orthodox inference (to adopt a fuzzy logic term) techniques can be chosen. In any case, the offline implementation of an equivalent look-up table has demonstrated to provide the best trade-off between complexity and quality of the results.

Overall, the proposed echo handling module provides, after the described delay estimation stage, a fine-tuned Geigel module, a custom AES module described above, possibly in combination with an adaptive NLMS module. Each of them

can work independently or in combination with the others, provided the necessary double talk information to the adaptive module. The complete echo handling block (and more generally, all the speech enhancement modules described in this paper) can be in this way tuned according to the available resources, echo characteristics and desired QoS; it constitutes a first version of general solution to the echo problem in portable devices that is ready to be adapted to hopefully improved characteristics of next generation devices and operating systems.

IV. Results and Field Tests

The audio quality enhancement toolset presented in this paper has been integrated in an existing VOIP softphone application, and tested on several different platforms; all the platforms are based on Windows Pocket PC and Smartphone with Windows CE 4.2 and 5.0, running on processors ranging from TI OMAP 850 with a clock speed of 195 MHz, to Samsung S3C at 400 MHz, to Intel Xscale devices at 400 and 600 MHz. Test conditions are characterized by typical noisy environment, hands-free mode and different device settings. The audio signal can be either at 8 kHz or 16 kHz.

In terms of CPU load, very different results can of course be observed on different devices and using different configurations and settings. On a TI OMAP based PDA, the complete processing path including active echo control without adaptive module can work for 8 kHz at a remarkable 25%; overall, 50% of the available resources are used by the complete softphone application during a normal call over wireless IP connection with good signal strength. On faster Xscale based platforms similar figures can be obtained operating in wideband for speech, possibly leaving room for sustainable low resolution video coding in case of narrowband speech.

In terms of quality, preliminary tests with two groups of speakers were conducted; the improved conversation quality was judged from "fair" to "good" in a normal MOS scale, when moving from simple signal processing and echo suppression to more advanced settings; conversation was considered clear and more intelligible with no exception (it was generally considered unintelligible and unacceptable just without echo control, whereas PLC and filtering can provide different levels of less categorical improvement, especially in function of the ear training level of the speakers). In single-talk situations, the AES module without the adaptive part was judged good. At extremely high levels of echo, suppression is not complete but still acceptable. In double-talk situations, the quality is clearly improved when the NLMS module is used, at the price of a doubling of the CPU time allocated to the echo control module.

V. Conclusion

In this paper we present a signal processing toolset for VOIP applications providing good quality audio enhancement and echo control especially targeting portable devices in hands free configuration. The toolset includes filters to reduce the perception of distortion, sampling-rate conversion, packet loss concealment and different echo handling modules allowing a flexible configuration. The toolset is successfully integrated into an existing softphone with more than acceptable results in devices running last generation processors and operating systems. A considerable improvement is still necessary in next generation platforms to permit the deployment of VOIP applications with a QoS for speech comparable or better than current PSTN telephone communications (many other services possible on IP are of course already available but cannot be directly compared).

Next steps in our work will include code optimization to allow improved, fully satisfactory quality in narrowband and wideband communication with reduced power consumption, as well as the integration of a more effective noise reduction for wideband and non-stationary noisy backgrounds.

References

[1] ITU-T Recommendation G.711 – Appendix I, "A high quality low-complexity algorithm for packet loss concealment with G.711", ITU-T, Geneva, Switzerland, 1999.

[2] Global IP Sound, "GIPS Enhanced G.711", White Paper, http://www.gips.com

[3] P. Dreiseitel, E. Hänsler, H. Puder, "Acoustic echo and noise control – A long lasting challenge," European Conference on Signal Processing EUSIPCO-98, Island of Rhodes, Greece, Conference Proceedings, vol. 2, pp. 945-952, Sept. 1998.

[4] J. Benesty, T. Gànsler, "A Multidelay Double-Talk Detector Combined with the MDF Adaptive Filter", EURASIP Journal on Applied Signal Processing, no. 2003:11, pag. 1056-1063.

[5] A. S. Chhetri, A. C. Surendran, J. W. Stokes, and J. C. Platt, "Regression-Based Residual Acoustic Echo Suppression," Proc. Intl. Works. on Acoust. Echo and Noise Control (IWAENC), Eindhoven, The Netherlands, September 2005.

[6] C. Faller, "Perceptually Motivated Low Complexity Acoustic Echo Control", 114th Audio Eng. Society Convention, Paper 5783, March 2003.

[7] C. Faller, C Tournery, "Estimating the delay and coloration effect of the acoustic echo path for low-complexity echo suppression", Proc. Intl. Works. on Acoust. Echo and Noise Control (IWAENC), Eindhoven, The Netherlands, September 2005.

[8] J.-S. Soo, K.K. Pang, "Multidelay block frequency domain adaptive filters", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 38, no. 2, pp. 373-376, 1990.

[9] D. L. Duttweiler, "A twelve-channel digital echo canceler," IEEE Transactions on Communications, vol. 26, pp. 647–653, May 1978.

[10] A. Mader, H. Puder, G. U. Schimdt, "Step-size control for acoustic echo cancellation filters – an overview", Elsevier Signal Processing Journal, vol. 80, pag. 1697-1719, 2000.