HANDSFREE SPEAKING
FOR COMMUNICATION TERMINALS

Hans J. Matt and Michael Walker
ALCATEL TELECOM, Lorenzstr. 10, D-70435 Stuttgart, Germany
Tel: +49-711-869-32246 and -32556; Fax: +39-711-869-32302
e-mail: hmatt@rcs.sel.de and mwalker@rcs.sel.de

ABSTRACT

In this paper some considerations for the realisation of a most *natural handsfree speaking* are presented. Its essential features comprise full duplex operation, speech loudness well adapted to the user's environment, background noise suppression and cancellation of line echoes. Furthermore its algorithms be able to work properly even under severe weaknesses caused by low cost components to allow the realisation of economic products.

1 INTRODUCTION

In real communication transmission one often faces situations as depicted in Fig.1. One person "A" wants to talk handsfree to another person "B" over a certain distance; local acoustic echoes occur on both sides as well as acoustic background noise. Line echoes may also occur due to hybrid circuits and line noise due to additive disturbances. Further to this the signals may be severely delayed during transmission.

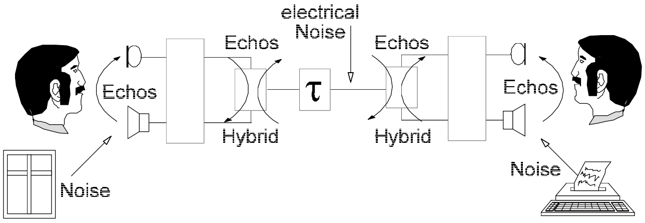

Fig. 1 Real transmission systems

Such a situation requests means for local acoustic echo suppression [1], line echo suppression, line attenuation control and noise reduction for the signals encountered. Further-more the specific acoustic environment on both sides needs to be matched.

2 ACOUSTIC ECHO REDUCTION

First we consider the acoustic path at user "A". He may speak from a distance of ≤1m to his telephone. Some background noise is present from office machines or people talking at distances > 4 m in the background. If his voice signal creates a microphone signal S_S which is greater than echoes from the loudspeaker S_L ($S_S/S_L \geq 0$ dB) or than background noise then our compander system alone eliminates echoes and noise with good performance [2-5]. In case there exists a stronger coupling between loudspeaker and microphone then further effort for echo reduction is needed.

2.1 Handsfree using a Compander

The principles of our handsfree speaking (Fig.2) are based on a compander system whose parameters are optimised with respect to psycho-acoustic masking effects [6].

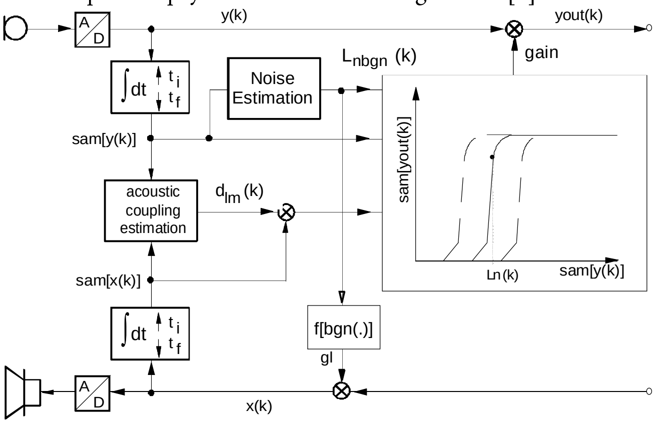

Fig. 2 Compander System

Voice signals at *normal loudness level* L_{no} are transmitted with just that loudness whereas louder signals are compressed to that normal level before transmission; signals at lower loudness are considered to be background noise and thus can be reduced further in level.

The key to realise this behaviour is the right estimation of signal levels [2-5], i.e. a *short time average magnitude* estimator ***sam(.)*** averages amplitude samples with an unsymmetric integration time constant, whereby increasing signal levels are tracked fast within a few ms and decreasing signal levels are followed in 60...100ms to avoid low frequency distortions. The $a_i(n)$ coefficients thereby create a weight function (in a time window) with an exponential decrease giving past values y(k) less influence.

$$(1) \quad \text{sam}\,[y(k)] = q_1 \cdot \sum_{n=0}^{\infty} [1 - a_i(n)] \cdot a_i(n)^n \cdot |y(k-n)|^p$$

with $a_i = a_r$ if $|y(k)| > \text{sam}[y(k)]$; else $a_i = a_f$
and preferably $p=1$; $a_r < a_f < 1$
q_1 = a normalisation factor.

Note that this function is in line with psycho acoustic rule since decreasing levels of $|y(k)|$ are masked by the preceding signals [6]. Due to its transfer characteristic the compander yields both, sufficient echo and background noise suppression.

2.1.1 Compander Adaptation to Background Noise

In a noisy environment people intuitively speak much louder to overcome ambient noise. In this case the value L_n for defining the *normal loudness* for the compander transfer function needs to be readjusted. This can be performed with the aid of a noise level detector. Noise level estimation can be done according to [7]; however we found and use a method with some lower computational effort. Within a continuously moving time window of a few seconds we then determine the minimum value of sam[y(k)] in that window as the estimate for the actual background noise bgn[y(k)],

$$(2) \quad bgn[y(k)] = \min \{ \, sam[y(k)] \, \} \quad \text{within} \quad t_u<t<t_o$$

and consequently adjust the value for $L_n(k)$ as

$$(3) \quad L_{nbgn}(k) = q_3 \cdot bgn[y(k)] \; ; \; q_3 = \text{const.}$$

$$(4) \quad L_n(k) = \max \{ L_{no} \; ; L_{nbgn}(k) \}$$

With this function we obtain an automatic adaptation of the compander transfer function to the ambient noise, i.e. in case of low noise $L_n(k)$ adapts to the fixed value L_{no} (transfer function staying in its most left position), whereas with increasing noise the transfer function is shifted continuously to the right.

2.1.2 Loudness Adaptation to Local Noise

Similar to the parallel shifting of the compander transfer function we adjust the loudness of the loudspeaker as a function of the local background noise. This is done via an adjustable gain factor *gl* in the path to the loudspeaker.

$$(5) \quad gl = f\,[bgn(.)]$$

Again this function matches psycho acoustic demands as the user in a noisy environment expects from his partner some louder speech.

2.1.3 Adaptation of Compander to the actual Coupling

The idea of parallel shifting of the compander transfer function in case of local noise can be further exploited in situations where a severe coupling between loudspeaker and microphone exists. Such a strong coupling may occur temporarily due to a sudden change in the room-acoustic characteristic or in situations where the design of a terminal inherently causes a strong coupling e.g. because the microphone is too close to the loudspeaker. In order to detect such strong coupling we decided to measure continuously the coupling value d_{lm} between loudspeaker and microphone and then adjust the compander transfer function to it. As a first rough estimate for d_{lm} we divided the *sam[y(k)]* by *sam[x(k-τ)]*. However this estimate appeared simply to be not good enough. Looking closer into the signal forms it became obvious that a better estimate for d_{lm} requires the precise determination of those time intervals where the far end speaker is significantly active.

Before we can select such time intervals we need to define (similar to eq.(1)) a *long time average magnitude* estimator ***lam(.)*** which averages amplitude samples with a symmetric integration time constant over >1 second.

$$(6) \quad lam\,[x(k)] = q_2 \cdot \sum_{n=0}^{\infty} b_i^{\,n} \cdot |x(k-n)| \;\; ; \text{ with } b_i<1$$

We then select the time intervals by the following rule:

i) First we look for a significant signal at the receiving path exceeding a certain threshold thr1, i.e.

(6a) $sam[\,x(k)\,] > thr1$ *and* $sam[\,x(k-\tau)\,] > thr1$

ii) Second we expect in addition a clear indication that the received signals x(k) are due to speech and not due to far end noise. This can be assumed if the sam(.) values exceed the lam(.) values for x(k); i.e. if both sam(.) > lam[x(k)].

iii) The minimum of the so found quotients is selected as the final d_{lm} from a continuously moving time window of size of a few seconds,

$$(7) \quad d_{lm}(k,\tau_1) = \min_{\substack{sam(x)>thr1 \\ and \\ sam(x)>lam(x)}} \{ \, sam[y(k)] \, / \, sam\,[x(k-\tau_1)] \, \}$$

whereby the parameter τ_1 is greater than the minimum delay τ between loudspeaker and microphone. This measurement technique allows to shift the compander transfer function smoothly to the right during speech from the far end. This gives additional adaptation to and stability in situations of critical coupling; however at the price of some audible reduction in openness during double talk situations. Thus a widely adaptive compander threshold can be realised by

$$(8) \quad L_n(k) \quad = \quad \max \{ \, L_{no}, L_{nbgn}(k), d_{lm}(k) \, . \, sam[x(k)] \, \}$$

2.2 Coupling from Loudspeaker to Microphone

For more critical loudspeaker to microphone arrangements - that appear in smart designed telephones - the coupling between loudspeaker and microphone yields higher and more critical values. One can observe then that echoes from the loudspeaker may create an up to ~25 dB higher signal than the user's voice (Fig. 1). This is not only difficult for echo cancellation, it influences the needed resolution of the A/D-converter following the microphone too. First the peak-value of the local echoes must not exceed the max. range of the A/D-converter to avoid non-linear distortions. Second for the user's signal at a~20 dB lower signal level there are just less quantization levels available (corresponding ~ 2...4 bit), which means the quantization noise of the A/D converter may become audible - depending on its total number of quantization levels.

2.3 Acoustic Cabinet Design

From the previous discussion it becomes obvious that the cabinet design of modern terminals for voice communication has a significant impact on the overall system performance:

i) In particular on the coupling from loudspeaker to microphone and consequently (in case of fixed A/D resolution) on the quantization noise and system margin against overload, which impacts harmonic distortions.

ii) The cabinet volume is responsible - together with the loudspeaker's lower corner frequency and resonances - for the achievable natural sound. It is known that the human ear perceives natural sound according to the rule

(9) $\sqrt{f_{lc} \cdot f_{hc}} \leq 900 \text{ Hz}$

i.e. one should not increase one corner frequency f_c without increasing the other according to eq. (9) as well.

iii) The components and materials used in the cabinet have a significant impact on its sound quality too; particularly in case of high loudness levels non harmonic distortions may occur due to cabinet resonances or loudspeaker overload.

For a known cabinet and loudspeaker frequency response we always use the possibility to compute and realise an equaliser filter in the path to the loudspeaker that flattens the frequency response in order to avoid significant peaks. However *smart designs* today still may cause some audible difference in sound quality if one compares the sound quality of usual terminals with that of small audio boxes.

2.4 Echo Reduction via Adaptive Filtering

In situations with strong coupling between loudspeaker and microphone (S_S/S_L< 0 dB) we need additional means to reduce the echoes to a level where the compander can work with much satisfaction. By theoretical considerations and experiments it was found that a combination of a short-length adaptive FIR filter together with the compander probably gives the best results. Some reasons for this are:

i) The FIR filter being localised directly after the microphone only needs to reduce the extremely strong coupling between loudspeaker and microphone in a small cabinet by some practical values in the order of ~10 ...25dB. This can be accomplished with a particularly short adaptive filter (e.g. using <80 coefficients) covering a range of <1..3m (depending from sampling rate) around the terminal.

ii) The compander located after the FIR filter then sees only a (reduced) coupling of S_S/S_L> 0 dB which he can handle with good performance. Further to this the FIR filter is somewhat critical in its behaviour, i.e. if its coefficients are instantaneously wrong adjusted it needs support from another side to keep the overall system stable. This important support can be provided by the compander too.

2.4.1 NLMS Algorithm for Adaptive FIR Filter

For calculation of the filter coefficients in real time we use the known *normalised least mean square algorithm (NLMS)*, which computes the actual coefficient $c_i(k)$ as

(10) $$c_i(k) = c_i(k-1) + \alpha \cdot \frac{y(k) \cdot x(k-\tau-i)}{\sum_{n=1}^{l} x^2(k-\tau-n)}$$

with l = length of the filter; τ = minimal delay between loudspeaker and microphone; α = const. ≤1 and

$$\sum_{n=0}^{\infty} x(.)^2 = \text{signal energy in the window.}$$

The NLMS algorithm was carefully analysed concerning its convergence properties; it converges only if

i) there is no local signal other than echoes being generated,

ii) there is a strong excitation signal x(k) from the receive path that owns enough bandwidth to generate uncoloured echo signals; e.g. a $\delta(t)$ function or noise would be good excitations, however sinusoidal excitations can easily misguide the algorithm.

iii) there are no non-linearities involved within the generation of echoes, since non-linear signal parts in the echoes can be interpreted as undesired additive local signals.

In practical situations these conditions are at least temporarily violated, e.g. during double talk or in the presence of other local signals. This causes the filter coefficients to get wrong adjusted which is equivalent to a divergence behaviour of the NLMS. Both convergence and divergence speed largely depend from the actual value of α.

2.4.2 Heuristic Filter Control

In view of these constraints we use the following rules:

a) The earlier explained condition for eq.(7) gives a first estimate for the excitation quality Q_e of signal x(k); we use this as a metric.

(11) $Q_e[x(k)] = \{sam[x] > thr2\} \cdot |sam[x] - lam[x]|$

with $Q_e[.] = 0$ if $sam[x] < thr2$, else $Q_e[.] \geq 0$

b) We further estimate the local speaker being active by

(12) $L_{spa} = sam(y) - \max\{L_{no}; L_{nbgn}; d_{lm} \cdot sam(x)\} \geq 0$

if $sam(y) > \max\{L_{no}; L_{nbgn}; d_{lm} \cdot sam(x)\}$

else $L_{spa} = 0$

c) As explained earlier we also estimate the local noise bgn[y(k)] according to eq.(2) and the coupling factor $d_{lm}(k)$ after the FIR filter according to eq.(7).

d) Further we measure the difference Q_{erle} between sam[y] after the microphone and sam[y1] after the FIR filter; if it becomes negative it indicates a loss by a misadjusted filter.

(13) $Q_{erle}(k) = sam[y] - sam[y1]$

e) Finally, since α determines the adaptation speed of the algorithm eq.(10) we then choose a function for α which heuristically increases or decreases the adaptation speed; i.e.

during time intervals of high confidence faster and during other time intervals much slower. In case of a negative loss it must increase α again to get fast out of the wrong status.

$$(14)\quad \alpha \sim \{a_1 \cdot Q_e^j(k) \cdot |d_{lm}(k) - \varepsilon_1|^p \cdot |1 - L_{spa} + \varepsilon_2|^m \cdot \left(\frac{bgn_{min}}{bgn}\right)^n + a_2\,[1 - \mathrm{sign}(Q_{erle})] \cdot Q_e^j(k)\ \}$$

The combined algorithm for the compander and short length adaptive filter is realised on a fixed point DSP; this became possible thanks to some *dynamic scaling* for the coefficients [3].

3. LINE ECHO REDUCTION

3.1 Far-End Echo Reduction

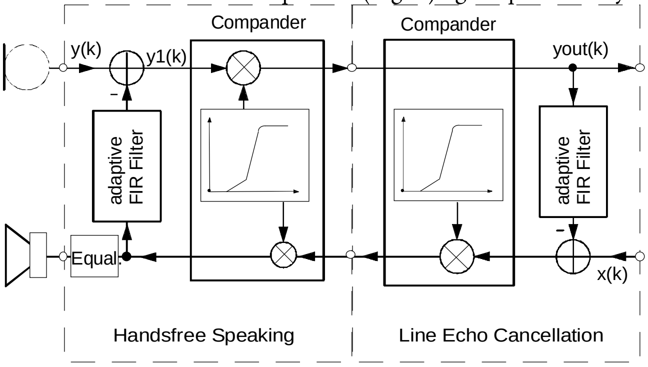

Fig. 3 Handsfree and Line Echo Reduction

With respect to line-echoes we face a different situation.
The far-end echoes may be generated at an arbitrary different location on the transmission path. They usually are smaller in amplitude than near-end echoes but little or nothing is known about the delay at which they occur. Furthermore the far-end speaker's signal amplitude and noise from the line are unknown.
For this situation the compander (Fig. 3) again proves very useful if he can be properly adjusted [4]. We therefore
i) estimate the speech level of the far-end speaker by

(15a) L_{fes}= lam [x(k)] only during time intervals t_i, where

(15b) $t_i = \{sam[x] > thr1\}\ \&\ \{L_{spa} = 0\}\ \&\ \{\tau_{spa}\ \}$; with

(15c) $\tau_{spa} = 1$ if $t > \tau_D$ (max. line echo delay time) after the event $\{L_{spa} = 0\}$; else $\tau_{spa} = 0$

ii) and adjust the line-echo compander according to

$$(16)\ L_{ln}(k) = \max\{L_{lno}, L_{lnbgn}[x(k)], d_{llm}(k) \cdot sam[y(k)], L_{fes}\}$$

3.2 Near End Echo Reduction

The near end echoes usually come from a near hybrid which produces almost strong but no time variant echoes; however these echoes can widely vary in amplitude from one connection to another one.

If the terminal must adapt to strong near-end echoes of ~(- 4dB) we use in addition to the compander a short-length adaptive linear filter (with <60 coefficients) to reduce the near-end echoes by ~10...15dB. The remaining near-end echoes after the filter again are then eliminated by the compander.

4. REALISATION OF TERMINALS

Broad band Handsfree System

Our first laboratory prototype operates with a fixed point DSP at a 24 kHz sampling rate. It is designed to meet the requirements of a true broad band i.e. a most natural communication at ~12 kHz. Its properties make it particularly suited for video-conferencing and for tele-cooperation applications. The algorithms presented so far were developed, analyzed and tested on it. Computational complexity is ~30 MIPS.

ISDN Videophone (G. 722)

The second realization of the handsfree is within Alcatel's videotelephone VKE S02 on a DSP operating with 16 kHz sampling rate. It meets the G.722 specifications of ITU and has a surprisingly good sound quality at 7 kHz bandwidth due to a good arrangement of loudspeaker and microphone. Computational effort is ~20 MIPS.

New Telephone Set

A new telephone set is currently under preparation in which we integrated our latest handsfree speaking version together with the line-echo cancellation part. It meets the G.711 specifications and has good sound quality under various acoustic environment conditions. Sampling rate is 8 kHz.

5. SUMMARY

In this paper we have shown how a good handsfree speaking system can be designed combining a compander with a short length adaptive FIR filter. It can be realized for several operational bandwidths and adapts automatically to different acoustic environments.

6. LITERATURE

[1] Hänsler E.: *"The handsfree telephone problem..."*. Annales des Télécommunications T 49 (1994), Nr.7-8, 360-367

[2] Walker M.:*"Handsfree speaking - a step towards a natural communication"*. Electrical Comm., 2nd Q.1993, pp. 181-187

[3] Walker M.; Matt H.:*"Verfahren zur Echokompensation"*. Deutsche Patentanmeldung DP 44 30 189.8

[4] Matt H.;Walker M.: *"Verfahren zur Kompensation von Leitungsechos"*. Deutsche Patentanmeldung DP 196 11 548.5

[5] Heitkämper P.; *"Freisprechen mit Verstärkungssteuerung und Echokompensation"*. VDI Fortschrittberichte Informatik/ Kommunikationstechnik Nr. 380; Düsseldorf 1995

[6] Zwicker E.:*"Psychoakustik"*. Springer-Verlag 1982, ISBN 3-540-11401-7

[7] Gierlich H.W.:*"Verfahren zur Sprachdetektion unter Störschalleinfluß"*. ITG-Fachbericht 105, 1988, S. 57-62