USING ORTHOGONALIZED VOICE FOR SIMULTANEOUS TRANSMISSION OF VOICE AND DATA

M. Goren, O. Tirosh, L. Kishon-Rabin and D. Wulich*
Department of Electrical & Computer Engineering,
Ben-Gurion University of the Negev.
Beer-Sheva 84105, POB 635, Israel.
Tel: ++972-7-461537, Fax: ++972-7-472949
e-mail: dov@bguee.bgu.ac.il

*School for Communication Disorders, Speech, Language and Hearing,
Sackler Faculty of Medicine, Tel-Aviv University, Israel

ABSTRACT

Voiceband channels are frequently used for data transmission, even though they were not designed for such a use. The reason is very simple; such channels already exist. It is also clear that such channels when used for data transmission can not be used at the same time for voice transmission, and vice versa. However, there are a lot of applications where simultaneous transmission of voice and data through the existing voiceband channel is needed.

In this work we propose a method for simultaneous transmission based on *orthogonalization* of the voice signal. A comprehensive assessment of the orthogonal voice which includes subjective measures shows that the orthogonal signal may have full intelligibility while its quality is only slightly degraded. The MOS for orthogonal voice is in the range 2.5 - 3.9 and depend on the data transmission parameters.

1. INTRODUCTION

There are a lot of applications where simultaneous transmission of voice and data through the existing voiceband channel is needed. Representative examples are: image telephone, simultaneous voice+fax, remote presentation, interactive games, etc., etc.

We show that it is possible to add, at the transmitter side, the voice signal to a digitally modulated data signal and to consider the voice signal as an interference. Using the fact that the voice signal is bounded (the minimal and maximal values are bounded) it is always possible to find a sufficiently low level of the voice signal that for a given modulation format will not cause any errors. Having the data detected without errors, it is possible to reconstruct the digitally modulated signal, to subtract it from the sum and so to obtain the voice signal. The above procedure can still be used when the channel background noise is very weak. In such a case it is still possible to obtain the Bit Error Rate (BER) at the needed value and to keep the Voice Signal to Noise Ratio (*VNR*) at an acceptable level [1].

The real problem arises when it is impossible to assume that the background noise is very weak. Here we propose a solution which is based on the idea described above, but we will *pre-process* the voice signal at the transmitter side, prior to its summation with the modulated signal. The pre-processing has to meet the following two requirements:
(1) the pre-processed voice signal will not cause any errors;
(2) the pre-processed voice signal will posses full intelligibility while the price paid for requirement (1) is only quality degradation.

The pre-processing proposed in this paper is based on the orthogonalization of the voice signal relative to the orthonormal basis of the modulated signal. The orthogonal voice signal, which is now orthogonal to the digitally modulated signal, due to this orthogonality, does not cause any errors; it only takes power from the total power budget of the channel. It will be shown that for certain parameters of the modulated signal such as the symbol rate and carrier frequency the orthogonal voice signal has full intelligibility and quite acceptable quality.

This paper is organized as follows. The principles of the solutions are given in Section 2. Section 3 is devoted to the orthogonalization idea. In this section the fundamental properties of the orthogonal voice signal and its effect on speech intelligibility are discussed. The experimental study results regarding the intelligibility and quality of various versions of the orthogonal voice signal are given in Section 4. Finally, in Section 5 the conclusion and discussion are given.

2. PRINCIPLES OF THE SOLUTION

Let us assume that the digitally modulated signal, which carries the data d_k is a passband PAM having the following form:

$$r(t) = \mathrm{Re}\left\{\sum_k d_k p(t-kT)e^{j\omega_c t}\right\}, \qquad (1)$$

where d_k represents the *k*-th symbol, $p(t)$ the shaping pulse, T the symbol duration and finally f_c the carrier frequency. We also assume that $p(t)=0$ for $t \notin [0,T)$.

Let $v(t)$ represent a voice signal which has to be sent simultaneously with the data. $n(t)$ denotes the background noise which is assumed to be bandlimited, white and Gaussian. It is also assumed that the voiceband channel is well defined by a bandpass filter (300-3400Hz).

The proposed solution to simultaneous transmission of voice and data is presented in Fig. 1. As shown in Fig. 1, the data, represented by a digitally modulated signal $r(t)$, is transmitted simultaneously with the orthogonal voice signal $v_R(t)$. Suppose for the moment, that the channel has constant transfer function over an unlimited bandwidth. Then the signal $v_R(t)+n(t)$ acts as an additive interference relative to $r(t)$. In fact $n(t)$ consists of the interference due to the orthogonality of $v_R(t)$ and is the only reason for bit errors. If we assume that the value of BER is very low, then almost full information about the signal $r(t)$ can be known at the receiver side. For a bandlimited channel the received signal is also influenced by the channel. This influence, however, can be tracked back by using adaptive techniques similar to those used in adaptive equalization. The channel influences both the orthogonal voice and the modulated signal $r(t)$. The digital demodulator has an equalizer of its self which compensate the channel influence such that the error rate is minimized. The demodulated data is re-modulated and the obtained signal is introduced to adaptive filter which produces at its output a signal $\hat{\hat{r}}(t)$ such that $E\left\{\left|\hat{\hat{r}}(t)-\tilde{r}(t)\right|^2\right\}\rightarrow \min$.

Assuming, that the voiceband channel is unknown, but can be well modeled as linear time invariant system, the optimal shape of the adaptive filter may be easily obtained by using a training sequence at the beginning of the transmission. (It may be the same sequence which is used for training the equalizer of the demodulator.)

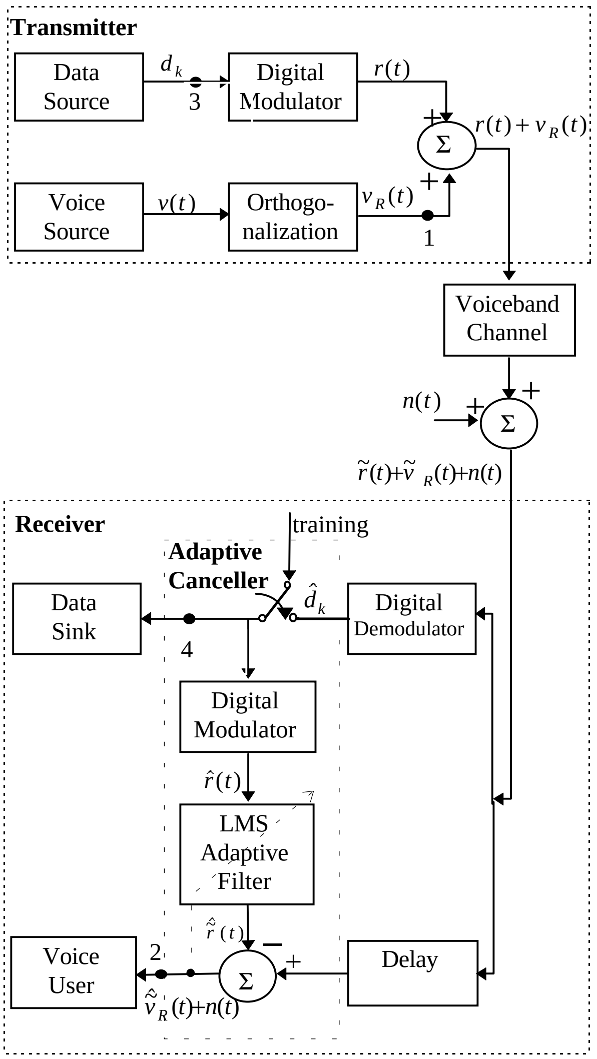

Fig. 1 Block diagram of the system for simultaneous transmission of the orthogonal voice and data.

Extensive computer simulation of such adaptive filter with the use of real voice signal shows that this method works very well for a “telephone line like” channel. An adaptive FIR filter of order 200 has been used [2]. It was obtained that for the ratio of modulated signal to orthogonal voice of +18dB at the input of the Adaptive Canceler and for BER of 10^{-3} (!) the ratio at the output (point "2") was -18dB, i.e., a rejection of the modulated signal of 36dB has been obtained in this case.

The influence of different BERs on the quality of the output signal has been checked in [3]. It was found that the quality is only mildly influenced by BER. For example MOS of 4.4 for BER=0 and clean speech signal (not orthogonalized) decreases to 4 for BER=$1.6\cdot 10^{-3}$.

The point "1c denotes the input and "2" denotes the output of the *equivalent analog channel* for transmission of $v(t)$ while the points "3" and "4" denote the input and output of the *equivalent digital channel* respectively.

3. ORTHOGONALIZATION OF $v(t)$

Let us, for real-valued $p(t)$, rewrite (1) for $kT \le t < (k+1)T$ in the following form:

$$\begin{aligned} r^{(k)}(t) &= \mathrm{Re}\{d_k\}p(t)\cos\omega_c t + \mathrm{Im}\{d_k\}p(t)\sin\omega_c t \\ &= \mathrm{Re}\{d_k\}\varphi_1(t) + \mathrm{Im}\{d_k\}\varphi_2(t) \end{aligned} \qquad (2)$$

where $\varphi_1(t) \overset{\Delta}{=} p(t)\cos\omega_c t$, $\varphi_2(t) \overset{\Delta}{=} -p(t)\sin\omega_c t$. $p(t)$ and ω_c are chosen such that $\varphi_1(t)$ and $\varphi_2(t)$ are orthonormal.

Using the Gram-Schmidt procedure let us find the *orthogonal voice signal* for $kT \le t < (k+1)T$:

$$v_R^{(k)}(t) = v^{(k)}(t) - a_k\varphi_1(t) - b_k\varphi_2(t), \qquad (3)$$

where

$$a_k \overset{\Delta}{=} \int_{kT}^{(k+1)T} v(t)\varphi_1(t)dt, \qquad b_k \overset{\Delta}{=} \int_{kT}^{(k+1)T} v(t)\varphi_2(t)dt \qquad (4)$$

and $v^{(k)}(t) = v(t)$ for $kT \le t < (k+1)T$. The above orthogonalization procedure can be illustrated by a phasor diagram - Fig. 2. The *orthogonal voice signal as a whole* is defined as:

$$v_R(t) \overset{\Delta}{=} \sum_k v_R^{(k)}(t). \qquad (5)$$

Introducing (3) into (5) yields

$$v_R(t) = v(t) - \sum_k \left[a_k\varphi_1(t-kT) + b_k\varphi_2(t-kT)\right]. \qquad (6)$$

The second term in (6) is the *in-plane* component of the voice signal. Let us denote it as $v_I(t)$. Consequently from (6) we have:

$$v_I(t) = \mathrm{Re}\left\{\sum_k c_k p(t-kT)e^{j\omega_c t}\right\}, \qquad (7)$$

where $c_k \overset{\Delta}{=} a_k + jb_k$. The *in-plane* component has a functional description similar to that used for passband PAM signals. Few restrictions, however are applied: (i) the coefficients c_k may obtain any complex value; they are not taken from a finite alphabet, (ii) the sequence $\{c_k\}$ is not necessarily (wide sense) stationary as a consequence of non-stationarity of the voice signal $v(t)$.

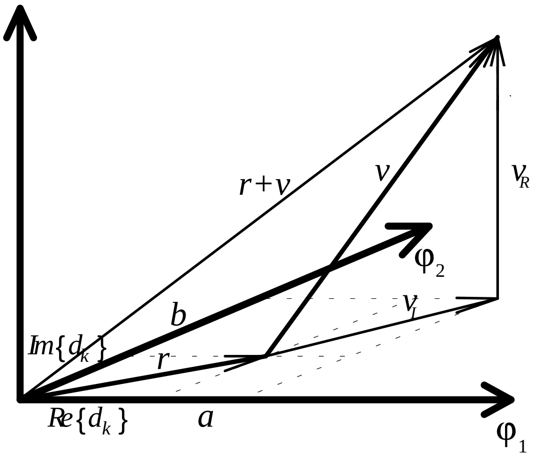

Fig. 2 The orthogonalization procedure - phasor diagram.

Transmitting $v_R(t)$ simultaneously with $r(t)$ will not influence the performance of a correlation-type digital demodulator. It takes only part of the transmitter's power. It is clear that $v_R(t)$ may be "far" from $v(t)$, according to the MSE criterion, for example. However, as shown in Section 4 the orthogonal voice signal $v_R(t)$ may be fully intelligible and can have good quality.

3.1 Properties of the Orthogonal Signal

As known [4], the voice signal may be considered as a wide sense stationary process on the time intervals of about 20 msec. Usually, the symbol duration T is much less than the stationarity interval of the speech signal T_s, i.e., at least for $k=1,2,...,\left[T_s/T\right]$ the coefficients a_k and b_k may be viewed, according to (4), as a response of the matched filter (correlation type digital receiver) to the stationary process. Consequently the coefficients a_k and b_k and therefore also c_k may be considered as a "short time" stationary sequence. This type of stationarity of a_k and b_k can be rigorously shown by direct inspection of $E\left\{a_{k_1}a_{k_2}\right\}$ and $E\left\{b_{k_1}b_{k_2}\right\}$.

Having $\left\{c_k\right\}$ stationary, the power spectral density of the *in-plane* component is given by a well known formula [5]:

$$S_{v_I v_I}(\omega)=\frac{1}{2}\left[U(\omega-\omega_c)+U^*(\omega+\omega_c)\right], \qquad (8a)$$

where

$$U(\omega)=\frac{1}{T}|P(\omega)|^2\sum_{m=-\infty}^{m=\infty}R_{cc}(m)e^{j\omega mT} \qquad (8b)$$

and $R_{cc}(m)$ is the autocorrelation function of the sequence $\left\{c_k\right\}$ and $P(\omega)$ is the Fourier transform of $p(t)$.

The results in (8) illustrate that the power spectral density of the *in-plane* component of the voice signal behaves, in its first approximation, similarly to the power density of the modulated signal $r(t)$ as given by (1), i.e., it is concentrated around ω_c and its main lobe has a width of $2\pi(2/T)$. In future work a more subtle analysis will be performed which will take into account the correlation properties of the sequence $\left\{c_k\right\}$.

It is clear from (6) that in order to make the orthogonal voice signal $v_R(t)$ as intelligible as possible it is sufficient to "move" the *in-plane* component as far as possible to the high frequency region of the channel by increasing $f_c=\omega_c/2\pi$ and to narrow its bandwidth by decreasing the symbol rate $(1/T)$: see eq (8).

4. EXPERIMENTAL RESULTS

Fifteen normal-hearing adults 20 to 36 years old, participated in tests in order to asses the hypotheses that:
(a) good speech intelligibility and quality is maintained after the orthogonalization process, and
(b) the intelligibility of orthogonalized signals are least affected when values of f_c are relatively high and values of $1/T$ are relatively small.

Quality was assessed using the *Mean Opinion Score* (MOS) test [6]. Intelligibility was assessed by the relative identification (in percent words correct) of Every-Day Sentence lists.

4.1 Speech Intelligibility Test

The group mean percent words recognized, and standard deviations for each of the four cases of orthogonal speech signals (processed conditions) are shown in Table 1. Although the scores were very high, a two-way repeated measures analysis of variance (ANOVA) on the arcsined-transformed data revealed a significant processing effect (F(3,42)=30.59, p<0.01). The processing conditions 3 and 4 did not have a negative effect on speech intelligibility. In fact, the results under these two conditions are similar to the mean score obtained by separate group of 15 normal-hearing subjects who listened to unprocessed speech (Table 1).

Processing Condition			**Word recognized**	
Case	**1/*T* [sps]**	***f*$_c$ [Hz]**	**Mean [%]**	**SD**
1	1600	1600	96.635	3.258
2	800	800	95.115	5.056
3	800	1600	99.791	0.554
4	800	2000	99.851	0.400
5	Unprocessed speech		99.783	1.026

Table 1. Mean percent words recognized, and standard deviation of the 15 subjects in each of the processing conditions and unprocessed speech. Each mean is based on word recognition of 30 every-day sentence list (i.e., 360 sentences).

4.2 Quality Rating Tests

MOS values were obtained for each of the processing conditions and are illustrated in Figure 3. It can be seen that values varied from 3.87 for the fourth processing condition ($1/T$=800sps and f_c=2000Hz) to 2.53 for the second processing condition ($1/T$=800sps and f_c=800Hz). The order of the processing conditions as found by MOS test are in perfect agreement with the hierarchy found using the first intelligibility assessment technique, that is, percent correct recognition of words sentences.

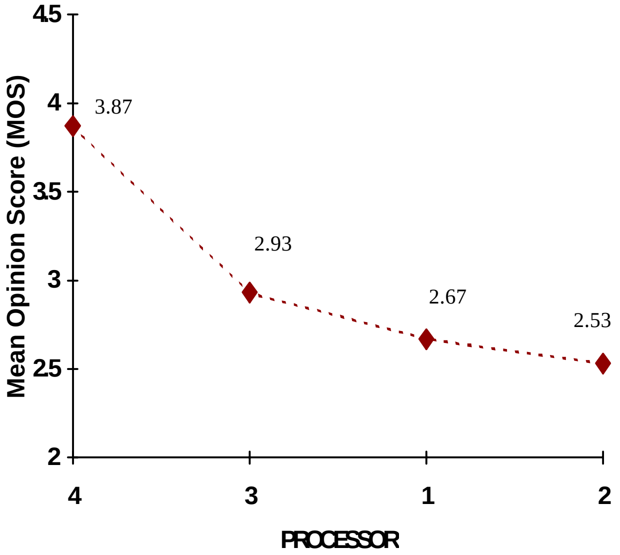

Fig. 3 MOS for various processing conditions.

5. CONCLUSIONS AND DISCUSSION

A method for simultaneous transmission based on *orthogonalization* of the voice signal is proposed. It is shown that it is possible to separate the orthogonal voice and data from the mixture of the orthogonal voice and digitally modulated signal, which implies that at the transmitter side the orthogonal voice and the modulated signal may be just added. Furthermore, the existence of the orthogonal voice in the received mixture does not cause any errors.

The quality and intelligibility of orthogonal voice depends on carrier frequency and symbol rate of the digitally modulated signal. Although results of the intelligibility test, as measured by percentage of words correctly identified, are in perfect agreement with the MOS values, intelligibility of every-day sentences is less affected by the orthogonal processing than its quality.

When a voiceband channel is considered (300Hz-3400Hz) the carrier frequency of 1600Hz and symbol rate of 800sps can be considered as an optimal choice. For such setting, as found, the intelligibility of the orthogonal speech are very close to that of an unprocessed speech. However, if the characteristic of the channel allows that, carrier frequency of 2000Hz will provide even better results (MOS of 3.87). The reason for that lies in the fact that frequencies upto 1500Hz are most important to speech perception. That is, important suprasegmental information and segmental information of vowels (first and second formants) lie in the low to mid frequency range.

The shape of the pulse $p(t)$ may also influence quality/intelligibility of the orthogonal speech. This will be examined in a future study.

Two parameters determine the performance of any method for simultaneous transmission of voice and data through a voiceband channel, namely the voice quality measured by MOS and the data rate R_d (for given probability of symbol error P_e).

The above mentioned parameters are being used for comparing the proposed method (ORT), to the three known methods: (i) Fully Digital (FD) [7]; (ii) Statistical Multiplexing (SM) [8], and (iii) Frequency Division Multiplexing (FDM).

For high *SNR* the MOS of the four considered methods is : MOS of ~3.7 for FD, MOS of ~3.8 for SM, MOS of ~3.9 for FDM and MOS of 2.53-3.87 for ORT.

The FD method can be applied where the *SNR* is such that the capacity of the channel is higher than the vocoder rate R_v. For example, for $R_v = 4.8\text{kb/s}$ the minimum *SNR* is about 3dB. For *SNR*s which are below such a minimum value the FD method can not be used while the ORT method may still be applied. The minimum value of *SNR* for ORT equals the value of *VNR* which assures the given a priori speech intelligibility and quality. From the above it can be concluded, that when the channel has a low *SNR* and there is a need to simultaneously transmit low rate data and voice, the ORT method may be successfully used while the FD method is useless.

6. REFERENCES

[1] L. Goldfeld, D. Wulich, "Performance of correlation demodulator in the presence of speech signal", *Signal Processing*, vol. 35, pp.41-50, 1994.

[2] M. Bukris, *Data Over Voice by Using Decision Feedback*, M.Sc. thesis, Dept. of Electrical and Computer Eng., Ben-Gurion University of the Negev, 1992.

[3] Y. Yosef, *Simultaneous Transmission and Detection of Data and Voice Using Decision Feedback*, M.Sc. thesis, Dept. of Electrical and Computer Eng., Ben-Gurion University of the Negev, 1993.

[4] N. S. Jayant, P. Noll, *Digital Coding of Waveforms*, Prentice Hall, Englewood Clifts, NJ, 1984.

[5] J.G. Proakis, *Digital Communications*, 2nd ed., McGraw-Hill, 1989.

[6] J.R. Deller, J.G. Proakis, J.H.L. Hensen, *Discrete-Time Processing of Speech Signal*, Maxwell Macmillan, 1993.

[7] *Digital Simultaneous Voice and Data*, Protocol Specification 1.2, Intel Corp., 1995.

[8] C. Roberge, J.P. Adoul, "Fast on-line speech/voiceband-data discrimination for statistical multiplexing of data with telephone conversations", *IEEE Trans. on Commun.*, vol. COM-34, pp. 744-751, 1986.