Variable-rate speech coding: Replacing unvoiced excitations by linear prediction residues of different phonemes

Wolfram Ehnert and Ulrich Heute

Institute for Network & System Theory, University of Kiel, Germany
Tel: +49 431 77572 401, Fax: +49 431 77572 403,
E-Mail: weh@techfak.uni-kiel.d400.de and uh@techfak.uni-kiel.d400.de

RÉSUMÉ

Afin de réduire le débit binaire de la transmission de la parole sans perte de qualité de celle-ci, nous développons un vocodeur qui utilise des méthodes differentes pour le codage des trames voisées et non voisées. Nous présentons ici une nouvelle idée de décrire des phonèmes fricatifs (sifflantes) et plosifs avec seulement 20 bit par trame de $t = 20ms$.

Nous montrons que ces phonèmes peuvent être représentés par des coefficients de la prédiction linéaire combinés avec un signal résiduel extrait d'un autre phonème prononcé par une personne differente connue à la station réceptrice du système de codage (voir figure 1). La présente contribution décrit aussi des algorithmes qui garantissent des transitions douces dans d'autres catégories de phonèmes.

En appliquant cette technique on peut considérablement réduire le débit de transmission (jusqu'à 1 kbit/seconde) pour les trames non voisées. Nous obtenons de meilleurs résultats qu'en utilisant des variantes de CELP (prédiction linéaire excitée par une table de codage) à 4 kbit/seconde. La combinaison de ce codage avec des méthodes de codage harmonique (par exemple le MBE: 'Multiband Excitation') pour les trames voisées resulte en un débit binaire variable de moins de 3 kbit/seconde.

ABSTRACT

In order to reduce the bit rate of speech transmission while maintaining the speech quality we are developing a vocoder which uses different methods for coding voiced and unvoiced frames. Within this framework we present an idea for expressing fricative and plosive phonemes with only 20 bits per frame ($t = 20ms$).

We show that they can be represented by LP(Linear Prediction) coefficients and a residual signal where this residue is always taken from a fixed phoneme of a test speaker known at the receiver station of the coding system (see figure 1). Algorithms ensuring smooth transitions to other speech-frame categories are also described below.

Using this technique the transmission rate of unvoiced frames can be considerably reduced (down to 1 kbit/s) getting better listening results than using CELP (Code Excited Linear Prediction) variants at 4 kbit/s instead. The resulting 'Multi-Class Vocoder' (voiced frames are coded by Harmonic Coding at 4 kbit/s) has a variable rate of less than 3 kbit/s on the average.

1 Introduction

Generally, speech can be subdivided into voiced, unvoiced and mixed (voiced/unvoiced) segments [1]. At bit rates of approximately 4 kbit/s, voiced speech can preferably be coded with Harmonic or Sinusoidal Coding (e.g. [2] and [3]) while the other speech categories miss a clear pitch frequency needed for the harmonic reconstruction of the short-time Fourier spectrum (STFS). Former approaches [4] [5] tried to code these frames (plosives, fricatives and zero frames) by Code-Excited Linear Prediction (CELP). The erroneous assumption [5] that the LPC-residues of plosives and fricatives are nearly white noise and can be coded only with stochastic codebooks resulted in rather poor speech quality for a bit rate of 4 kbit/s. The introduction of adaptive codebooks, however, requires a higher bitrate. Furthermore, CELP coding realizes a concatenation of artificial residual vectors each one being chosen according to criteria based on the comparison of the synthetic result to the original frame. It is evident that a non-split excitation of a real human voice represents the physical speech production more accurate, thus leading to better coding results for unvoiced frames (at this bit rate). In the following two paragraphs we show how such residues can be used instead of the well-known residual-vector codebooks of CELP. In the succeeding section we discuss possible realizations of a variable-rate coder based on this idea.

2 Replacing mechanism, main idea

The original samples $s(k)$ ($k = 1..S$, S = number of samples per frame) of each speech frame can be represented by the LPC coefficients a_i ($i = 1..p$, p = filter order of the linear predictor) and the residual samples $r(k)$ according to

$$s(k) = r(k) - \sum_{i=1}^{p} a_i s(k-i). \qquad (1)$$

In an ideal coding system both $r(k)$ and a_i will be calculated at the transmitter station and transmitted totally to the receiver where the original signal $s(k)$ can be recovered. For lack of channel capacity in real applications, the transmitted data must be reduced as much as possible. The LPC coefficients might be coded as Line-Spectral Pair (LSP) frequencies with non-linear quantizers or vector codebooks. Following the idea of replacing $r(k)$ by residues of different phonemes we now can cancel $r(k)$ totally (see figure 1).

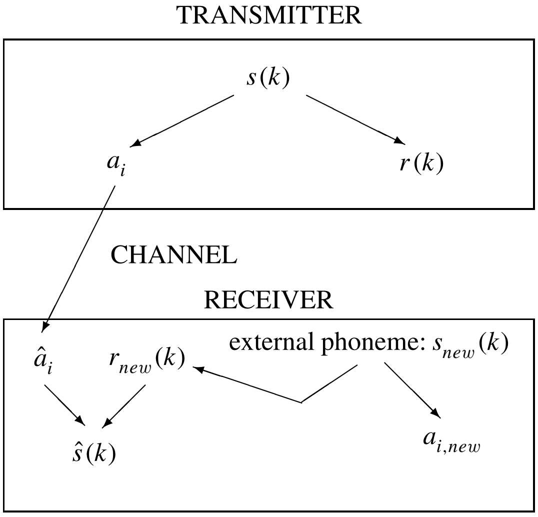

Figure 1: Coding scheme of fricatives and plosives

At the receiver station we obtain a synthesized signal $\hat{s}(k)$ by filtering the speaker-independent residue $r_{new}(k)$ with the decoded LPC coefficients $\hat{a}_i$. As the experimental results show, there is no need to provide different $r_{new}(k)$ for each kind of phoneme. It is sufficient to distinguish between fricatives and plosives. The residue $r_{new/fric}(k)$ of the phoneme 'sh' filtered with the received coefficients $\hat{a}_i$ of any other fricative phoneme (e.g. 's', 'f' etc.) turns into the fricative sound corresponding to $\hat{a}_i$. This corresponds to the well-known fact that a sound is mainly described by its spectral envelope rather than the fine structure. In the same way it is possible to generate any kind of unvoiced plosives (e.g. 'k', 't' etc.; note that voiced plosives like 'b' should be processed with harmonic coding) out of the corresponding LPC coefficients $\hat{a}_i$ and the residue $r_{new/plos}(k)$ of a plosive phoneme of a different speaker.

According to the experiments, in both cases $r_{new/fric}(k)$ and $r_{new/plos}(k)$, this external speaker should have preferably a female voice with a higher pitch frequency (because the spectrum of the residue of a female voice is supposed to be more constant at frequencies close to $f = 4kHz$). If possible, the fricative substitution residue should be longer than the fricative phoneme to be coded. Thus, the LPC coefficients of each speech frame can be passed through the channel in order to filter always the next S residual samples $r_{new/fric}(k)$ until the original phoneme has finished. The transition to the following zero frame or voiced speech segment can be realized with adequate windowing. Transitions from fricatives to plosives do not exist because all plosives are preceded by zeros [6]. This is one of the reasons why the synthesis of plosive frames can be realized in a different and even more efficient way as shown in the next section.

The residues $r_{new/fric}(k)$ and $r_{new/plos}(k)$ are normalized vectors. Therefore, we have to transmit the energy factor (gain factor) as well. However, the real advantage of the residue replacement still consists in not transmitting the residues.

3 Replacing plosive residues

Plosive phonemes (unvoiced stops) are always produced by an interruption of the airstream through the vocal tract (occlusion) followed by an air burst [6]. Thus, the energy containing samples of a plosive are always preceded by zeros.

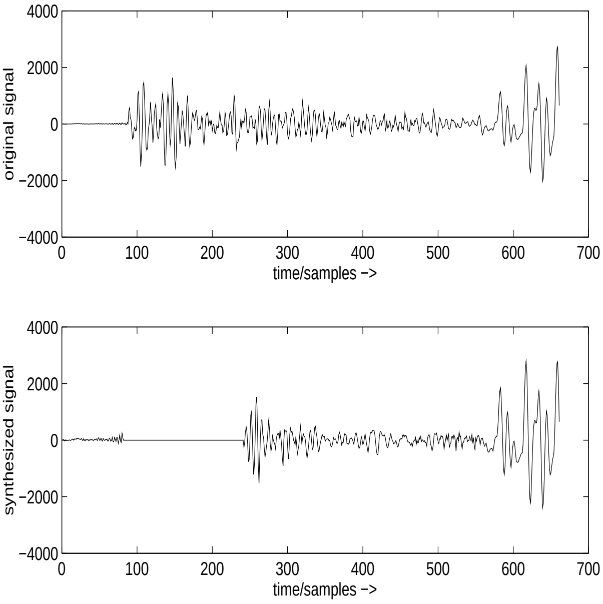

Figure 2: Original and synthesized phoneme 'k' of 'cases'

This zero segment can be extended easily by adding new zero samples. Therefore, we can define the beginning of the air burst by ourselves. The actual realization of the replacement algorithm turns each plosive phoneme of any length into a phoneme (out of $r_{new/plos}(k)$ and $\hat{a}_i$) with the fixed length of exactly two frames. Figure 2 compares exemplarily the synthesized signal to the original samples (the listening impressions of this result will be discussed in the following sections).

If we permit a delay of two frames (40ms) at both transmitter and receiver, then it is now even possible to generate the plosive phoneme in the same way as described for the fricative phoneme but with a trick:

Figure 3 shows a possible sequence of frame classifications for the original speech: Three plosive frames (representing one plosive phoneme) are followed by a voiced frame. The synthesis procedure always knows two frame classifications in advance.

At the moment $T = 0$ the decoding system already received the decision of point $T = 2$ to be a plosive frame. As the plosive which replaces the original one is only two frames long, it is now obvious that it cannot start at $T = 0$. As mentioned above, plosives are always preceded by zeros. Therefore, the frame at $T = 0$ can be filled up with zeros. At $T = 1$ the receiver will be informed about the frame of $T = 3$ which is voiced. Now the substitution algorithm can start using the plosive residue of the examplary phoneme:

The synthetic signal will be constructed backwards by calculating the first values of the voiced frame using its harmonic information (spectral magnitudes, pitch frequency etc.). These samples (marked with an arrow in figure 3) serve (after being flipped in time from forward to backward) as initial values for the Linear Prediction filter with the coefficients received for the respective plosive frame. $r_{new/plos}(k)$ must now be passed through this filter starting with $k = 320$ down to $k = 1$. Thus, we get a smooth transition between the plosive and the following voiced frame. Transitions between plosives and fricatives are handled in the normal way using the last plosive samples as initial values for the succeeding fricative LP filtering.

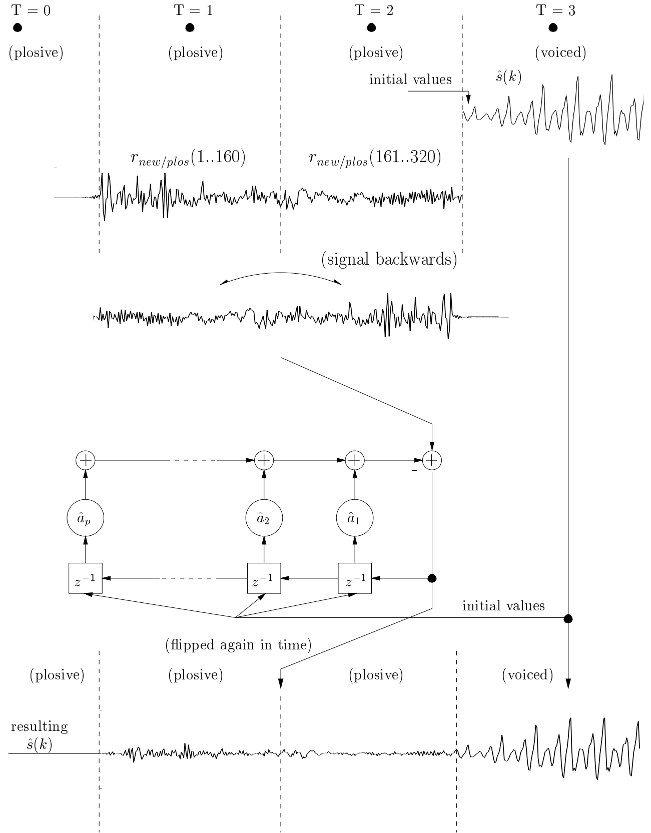

Figure 3: Reconstruction of plosive phonemes

$r_{new/plos}(k)$ is the residue of a natural plosive phoneme. This phoneme has once been cut at the transition to a voiced frame. Therefore, it was necessary to find a new transition algorithm. The transition from the preceding zero samples (interruption of the airstream) to the plosive, however, has been included in the phoneme. Therefore, this transition is expressed by the residue $r_{new/plos}(k)$ itself and needs no correction (smoothing algorithm) anymore.

4 A coder based on phonetic categories

The resulting coder using the residue-replacing algorithms presented above is a variable-rate coder because voiced and unvoiced speech frames are coded with different methods and rates. The resulting 'Multi-Class Vocoder' needs to distinguish between voiced frames, plosives, fricatives, and zero frames. If we already allow a delay of 40ms we know the original samples of the two succeeding frames. Then it is possible to obtain a quite acceptable (but never perfect) classification of non-disturbed speech. Mainly three errors can occur:

– A voiced frame is interpreted as a fricative or plosive: In this (worst) case the periodic signal will be replaced by the LP-filtered fricative or plosive residue. Since the residue does not contain any pitch frequency, the synthesized frame is aperiodic as well. Furthermore, it is impossible to perform an acceptable transition to previous or next frames which might be voiced and correctly recognized as such. These cases as well as the next one fortunately happen rarely.

– An unvoiced frame is interpreted as a voiced segment: The effect is known from the first harmonic-coding approaches. The synthesized signal sounds metallic because of the missing random components of unvoiced speech segments. The lack of naturalness could be reduced using a Multiband-Excitation (MBE) vocoder [2] instead of Harmonic Coding for coding voiced frames (see following section).

– Fricatives are coded as plosives and vice versa: This confusion can be explained by the fact that some speech segments are phonetically hard to define: A 't' can be spoken in different ways and might sound more fricative than plosive. In this special case the plosive component of the fricative-sounding 't' will be generated automatically by the transition from the preceding zero frame to the synthesized fricative. The synthesized 't' is phonetically closer to the original speech when using $r_{new/fric}(k)$. It results in an advantage that the classificator attaches more importance to the sound than to the spelling of words.

Figure 4 shows the structure of the introduced vocoder. The results are far better than those using a CELP coder (at 4 kbit/s) for unvoiced frames [5]. The bit rate, however, is going down to less than 3 kbit/s if we code the fricatives and plosives with 20 bits per frame (20ms). Tests have even shown that the LPC coefficients of fricatives can be updated for only each second frame without changing the speech quality.

Till now we achieved the best results combining the 4.15 kbit/s - IMBE vocoder (Inmarsat-M standard) with the new substitution algorithm for unvoiced phonemes as presented above. In some cases the fricatives and plosives coded by the IMBE vocoder itself (applied for voiced and unvoiced frames) resulted to be closer to the original phoneme. In other cases especially the fricatives sounded reverberant in the IMBE version and could be highly improved by the substitution algorithm. In total, improvements and deteriorations of the listening results of the new 'Multi-Class' (MC) vocoder seem to keep the balance in comparison to the former IMBE realization of Griffin and Lim [2] which did not distinguish methodically between voiced and unvoiced frames. However, at the moment the actual 'Multi-Class' version including the recognition of the mentioned phoneme categories is far away from being realized in real time. Furthermore, the delay of $t_d = 40ms$ at both transmitter and receiver may exceed the demands and standards of a telefonic system. There are ideas to reduce the delay of the receiver to only one frame ($t = 20ms$) using a different algorithm for the transitions from plosives to voiced frames. In some cases, however, this time saving might result in a worse phoneme detection.

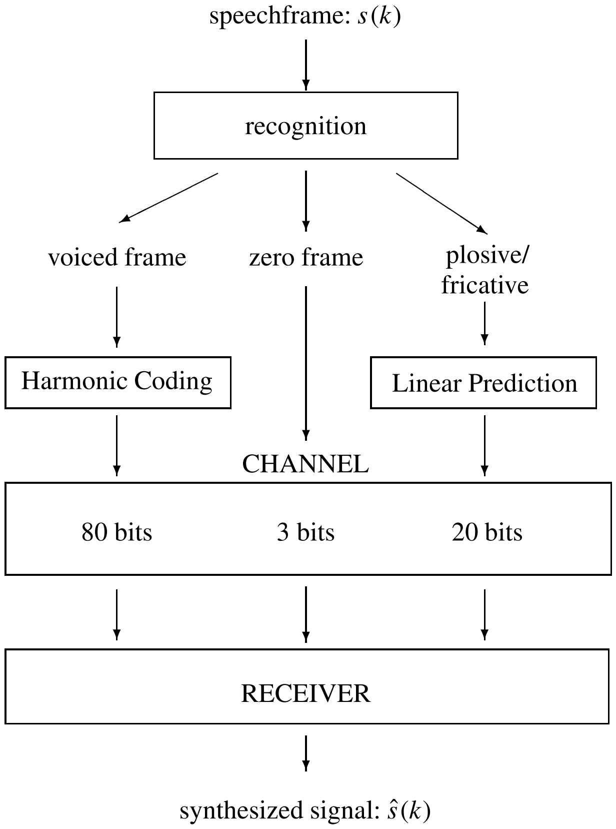

Figure 4: Proposal for a variable-rate coder

5 Applications

As described in the section above it is still not possible to use the new substitution algorithm for a real time realization of a coding system. TTS (Text-to-Speech) systems, however, need at best the synthesis algorithm to be realized in real time (all other delays are not relevant). The coding will be applied only once (before application) in order to get a memory-efficient database of phonemes, diphones or syllables. If a phoneme of the original speech data has not been recognized correctly, the classification can be rectified 'by hand'. As the new 'Multi-Class' version is already concatenating voiced and unvoiced speech parts it might be useful for a concatenation-based TTS system. At present, investigations are being carried out on this topic.

6 Conclusion

Residues of fricatives and plosives do not represent white noise. Nevertheless, they can be replaced by different residues without changing the individual characteristics of a speaker, thus getting better listening results than by using the residues of a stochastic codebook. For the coding application unvoiced frames must be subdivided into plosives, fricatives and zero frames, but a confusion of plosives and fricatives does not result in a wrong phoneme at the receiver station. Unvoiced frames are represented by only one set of LPC coefficients and the gain. Therefore, we get a highly reduced variable rate (below 3 kbit/s on the average) for the Multi-Class Vocoder which is, however, not realizable in real time up to now.

References

[1] A. Das and A. Gersho, "Variable dimension spectral coding of speech at 2400bps and below with phonetic classification," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 492–495, May 1995.

[2] D. Griffin and J. Lim, "Multiband excitation vocoder," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 1223–1235, August 1988.

[3] R. McAulay and T. Quatieri, "Speech analysis-synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 744–754, 1986.

[4] I. Trancoso, L. Almeida, and J. Tribolet, "A study on the relationships between stochastic and harmonic coding," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1986.

[5] S. Otte, *Mehrklassen-Sprachkodierung: Verarbeitung periodischer Rahmen*. Diploma, LNS/CAU Kiel, 1996.

[6] M. Torres and P. Iparraguirre, "Acoustic parameters for place of articulation identification and classification of spanish unvoiced stops," *Speech Communication*, vol. 18, pp. 369–379, 1996.