PERCEPTUAL LINEAR PREDICTIVE NOISE MODELLING FOR SINUSOID-PLUS-NOISE AUDIO CODING

Richard C. Hendriks, Richard Heusdens and Jesper Jensen

Dept. of Mediamatics
Delft University of Technology
2628 CD Delft, The Netherlands
email: {R.C.Hendriks, R.Heusdens, J.Jensen}@EWI.TUDelft.nl

ABSTRACT

Sinusoidal coding of audio subject to a bit-rate constraint will in general result in a noise-like residual signal. This residual signal is of high perceptual importance; reconstruction of audio using the sinusoidal representation only will typically result in an artificial sounding reconstruction. In this paper we present a method, called perceptual linear predictive coding (PLPC), where the residual is encoded by applying LPC in the perceptual domain. This method minimizes a perceptual modelling error and therefore represents only residual components that are of perceptual relevance, while automatically discarding components masked by the sinusoidal coded part. Subjective listening tests show that PLPC performs significantly better than ordinary LPC as a sinusoidal residual coding technique. Furthermore, PLPC combined with a flexible segmentation and model order allocation algorithm leads to a significant gain in terms of R/D performance for fragments with fast changing characteristics.

1. INTRODUCTION

Sinusoidal coding has proven to be an efficient technique for low bit-rate coding of audio, see e.g. [1]. Here, a signal is represented by a sum of sinusoids, where parameters as amplitudes, phases and frequencies are extracted from the underlying signal. The number of sinusoids that are extracted is in general limited due to bit-rate constraints. Consequently, broad-band noise-like components are typically not represented by sinusoids, since the required number of sinusoids would be too large. Instead, a separate coder is used to encode the noise part of the signal. The sinusoidal coder often uses a perceptual estimation criterion, for example a perceptual norm [2], to extract the tonal components. The residual signal that results after extraction of sinusoids is then modelled by the noise coder.

One common approach of residual coding is to use a filter bank based on the human auditory system [3]. In this approach the spectrum of the residual signal is divided into equivalent rectangular bandwidths (ERBs), where the energy level in each of the bands is computed, quantized, encoded and transmitted to the decoder. A drawback of this method is the relative high bit-rate needed to encode those parameters (typically 8 kbit/s, see [4]).

A complication of coding the residual signal is that high-energy but perceptually unimportant tonal components that are still present in the residual signal may be represented by the noise coder, rather than the perceptually important noise-like components. In that case bits are wasted on parts of the spectrum that are masked by the sinusoidal reconstruction.

One way to overcome the above mentioned problem is to use a method as described in [5] where perceptual irrelevant sinusoids are removed from the target signal using an iterative procedure, prior to noise coding. The iterative procedure is terminated based on a heuristically determined criterion.

In this paper we present an alternative, efficient method to encode the residual after sinusoidal modelling that overcomes the above mentioned complication without making use of heuristic criteria. The method is called perceptual LPC (PLPC). It is based on LPC in the perceptual domain, leading to encoding of the perceptual important components of the residual only. An advantage of using LPC is that the technique itself is well understood and that efficient quantization and encoding techniques can be adopted from e.g. the speech coding field [6]. To avoid confusion with the LPC related term 'LPC-residual', we will refer from now on to the residual signal after sinusoidal modelling as the 'target signal'. We note that Hermansky used the name perceptual LPC in [7] for an equal loudness curve based LPC technique in the context of speech recognition. This method, however, does not take into account spectral masking.

This paper is organized as follows. In Section 2 the concept of PLPC will be explained and will be validated with an example showing the advantages of perceptual LPC over ordinary LPC. Section 3 discusses two practical aspects of the proposed method, while Section 4 explains the use of variable-length time segmentation and model order allocation for PLPC. In Section 5 some experimental results are discussed and finally, in Section 6, some conclusions will be drawn.

2. RESIDUAL CODING WITH PERCEPTUAL LPC

LPC coefficients are normally found by minimization of the energy, or l_2-norm, of the modelling error. It can be shown [8] that this l_2-norm can be rewritten as

$$E = \int_0^1 \frac{P(f)}{\tilde{P}(f)} df, \tag{1}$$

where $P(f)$ and $\tilde{P}(f)$ are the original and LPC-model power spectrum, respectively. A consequence of (1) is that the approximation of $P(f)$ by $\tilde{P}(f)$ is better at the spectral peaks, where more energy is present than in the valleys of the spectrum [8]. The fact that LPC tends to model spectral peaks gives rise to modelling problems

when the target signal is a residual signal where the perceptual important sinusoidal components are already extracted. This target signal does not only contain perceptually important noise components, but perceptually unimportant (sinusoidal) components as well. Although those sinusoidal components will be masked by the perceptual relevant sinusoids, they can contain more energy than the perceptually more important noise-like components in the target signal and may, therefore, dominate the power spectrum. Modelling of the power spectrum $P(f)$ by $\tilde{P}(f)$ will then lead to inaccurate modelling of perceptual important noise-like parts of the spectrum. Rather than minimizing the well known l_2 distortion measure as in (1), we minimize in this paper a perceptual relevant distortion measure. This new approach has the advantage that no preprocessing step is needed to remove the remaining perceptually irrelevant tonal components as in [5], since the perceptual distortion measure will automatically recognize that those components have no perceptual relevance and will, therefore, not be modelled. The perceptual distortion measure we use is defined in [9] as

$$\|\varepsilon\|_\pi^2 = \int_0^1 \hat{a}(f)|\hat{\varepsilon}(f)|^2 df, \tag{2}$$

where ε is the modelling error, $\hat{}$ indicates the Fourier transform operation and $\hat{a}$ is the reciprocal of the masking threshold. The relation between ε and the original signal y is given by [10]

$$\varepsilon = y - \sum_{k=1}^{p} \alpha_k y(\cdot - k),$$

with α_k the LP-coefficients and p the model order. From [9] it follows that $\hat{a}$ is positive and real for all $f \in [0, 1)$, and therefore (2) defines a (perceptual) norm, which we will refer to as $\|\cdot\|_\pi$. In order to only spend bits in spectral regions of perceptual importance, we find the LP-filter coefficients α_k which minimize the perceptual norm (2), rather than minimizing the l_2 norm of the modelling error which is done in standard LPC. We can rewrite our minimization problem as:

$$\begin{aligned}\min_\alpha \|\varepsilon\|_\pi^2 &= \min_\alpha \int_0^1 \hat{a}(f)|\hat{\varepsilon}(f)|^2 df \\ &= \min_\alpha \int_0^1 |\hat{h}(f)\hat{\varepsilon}(f)|^2 df \\ &= \min_\alpha \|(h * \varepsilon)\|_2^2, \end{aligned} \tag{3}$$

leading to a minimization of the l_2-norm of a filtered modelling error, where the filter $\hat{h} = \hat{a}^{\frac{1}{2}}$. An interpretation of (3) is that the convolution between h and ε defines a transformation of ε to a modelling error ε_π in the perceptual domain, that is, $\varepsilon_\pi = h * \varepsilon$. From (3) it thus follows that minimizing the perceptual norm of the modelling error ε is equivalent to minimizing the l_2-norm of the perceptual modelling error ε_π, that is

$$\min_\alpha \|\varepsilon\|_\pi^2 = \min_\alpha \|\varepsilon_\pi\|_2^2. \tag{4}$$

Thus, to minimize the perceptual norm of the modelling error, we can apply LPC to the signal transformed to the perceptual domain.

Figure 1 shows a block diagram (both encoder and decoder) of the PLPC scheme. The target signal $y_{\rm res}$ is first filtered by a filter $\hat{h} = \hat{a}^{\frac{1}{2}}$ which is determined by the psychoacoustical model (PA model). The function $\hat{a}$, and thus the masking threshold, is computed on the basis of the reconstructed sinusoidal signal $\tilde{y}_{\rm sin}$

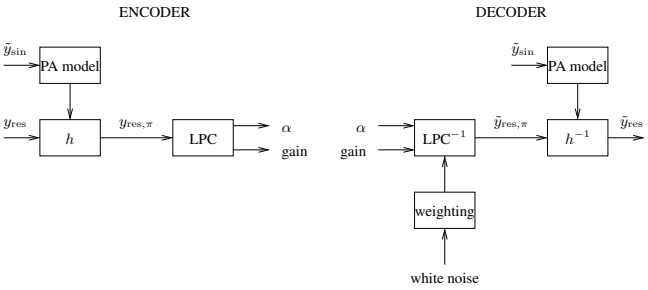

Fig. 1. *Scheme (both encoder and decoder) of perceptual LPC.*

only. This has the advantage that no side information has to be transmitted since $\tilde{y}_{\rm sin}$ is known at the decoder. The filtered signal, denoted by $y_{{\rm res},\pi}$, is analyzed in the block labeled LPC where the LP-filter coefficients α and a gain factor are computed. At the decoder side, the reconstructed signal $\tilde{y}_{{\rm res},\pi}$ is generated by filtering a colored noise signal (to be explained in Section 3) using the received filter coefficients α and gain factor. Finally, $\tilde{y}_{{\rm res},\pi}$ is filtered by h^{-1}, resulting in the reconstructed target signal $\tilde{y}_{\rm res}$.

In order to demonstrate the advantages of PLPC over ordinary LPC, we applied both methods to a signal consisting of two sinusoids located at 1300 and 1350 Hz in additive second-order autoregressive noise. The spectrum of this signal is shown in Figure 2a. Because the amplitude of the sinusoid at 1300 Hz is eight times as large as the amplitude of the sinusoid at 1350 Hz, the 1300 Hz sinusoid will cause (at least partly) spectral masking of the 1350 Hz sinusoid. With a sinusoidal estimation algorithm based on the perceptual distortion model as described in [9], the perceptually most relevant sinusoid is extracted from the original signal. This results in selection of the 1300 Hz sinusoid, of which the magnitude spectrum is shown in Figure 2b. The magnitude spectrum of the resulting residual signal containing the perceptually less relevant sinusoid at 1350 Hz together with the autoregressive noise is shown in Figure 2c. This target signal is encoded in two different ways: by applying ordinary LPC and by applying PLPC. Applying ordinary second-order LPC on this target signal will result in a LP-filter of which the magnitude response is depicted in Figure 2e. Clearly the LP-filter is biased towards the tonal component, while the perceptually much more important noise component is not modelled accurately. Synthesis of the target signal and adding it to the sinusoidal coded part leads to the total reconstruction, of which the spectrum is depicted in Figure 2g. Here, the second-order autoregressive noise is completely missing, resulting in a perceptually degraded reconstruction.

This problem can be overcome by usage of PLPC as defined by (4). To do so, the target signal is transformed to the perceptual domain by h, which results in the spectrum shown in Figure 2d. In comparison to the noise, the sinusoid at 1350 Hz is now suppressed which is in line with the spectral masking caused by the perceptually more important sinusoid at 1300 Hz. Applying second-order LPC to the transformed signal will result in the LP-synthesis filter whose magnitude response is shown in Figure 2f. By usage of this LP-filter the target signal can be reconstructed in the perceptual domain and back-transformed by h^{-1} to the l_2-domain. Figure 2h shows the final reconstruction in the l_2-domain consisting of the modelled target signal added to the sinusoidal coded part. As expected, Figure 2h shows accurate modelling of the perceptually important AR noise and less accurate modelling of the perceptual less important 1350 Hz sinusoid.

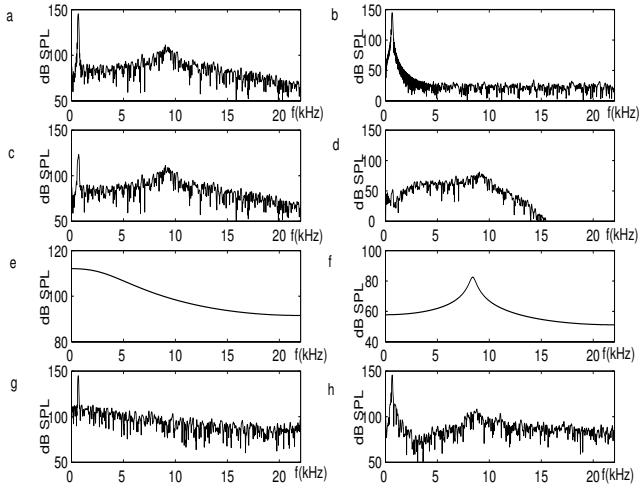

Fig. 2. *Example showing the difference between LPC and PLPC.*

By comparison of the reconstructions by LPC and PLPC, two conclusions can be drawn. In the case where LPC is used in the l_2 domain (see Figure 2g), a broad spectral region around the sinusoid located at 1300 Hz is severely degraded by noise. Secondly, the AR noise present in the original signal is not modelled at all. With PLPC (see Figure 2h), the AR noise component is reconstructed accurately.

3. PRACTICAL ASPECTS

As mentioned, we need information about the masking threshold, and thus $\hat{h}$, at the decoder. To avoid bit-rate expensive transmission of h^{-1} to the decoder, we assume that masking curves are dominated by the perceptual important tonal components of the original audio signal. With this assumption, the perceptual inverse transformation h^{-1} can be determined unambiguously from the sinusoidal coded part and comes therefore for free in terms of bits, since that information is already transmitted to the decoder. The validity of this assumption will be shown experimentally in Section 5.

A second practical aspect is concerning the perceptual modelling error. It will, in contrast to ordinary LPC, not be whitened over the whole frequency range. The reason for this is that the perceptual transformation is dependent on the reciprocal of the masking threshold in quiet, which has a very fast decay at very low and high-frequency spectral regions. After transformation to the perceptual domain, the target signal is dominated at low and high frequencies by this fast decay. Unfortunately, these steep spectral slopes cannot be modelled well with a (low-order) LPC model, and, consequently, PLPC modelling errors become non-white. To be able to still use Gaussian substitutes for the modelling error of the PLPC process, we use white excitation signals that are shaped by the masking threshold in quiet at very high and low frequencies. This spectral shaping of the PLPC excitation signal is done in the block labeled 'weighting' in Figure 1.

4. VARIABLE LENGTH SIGNAL ANALYSIS

In order to better adapt to the local statistics of the signal and to avoid quantization and modeling errors in front of signal transients (a phenomenon which is known as pre-echoes), we perform a variable-length signal analysis where the LPC model order varies over segments. To do so, we use the algorithm described in [11], where we compute distortions using the perceptual distortion measure (2). The algorithm finds an optimal segmentation $s = \{s_1, \ldots, s_N\}$ consisting of N variable-length segments, each having a length which is an integer multiple of a predefined minimum segment length (in our case 2.9 ms). In addition, it finds the set of optimal model orders $p = \{p_{s_1}, \ldots, p_{s_N}\}$, where p_{s_k} denotes the optimal model order for segment k. By optimal we mean optimal in terms of rate and distortion (see [11] for details).

5. EXPERIMENTAL RESULTS

We implemented several versions of the proposed method for sinusoidal residual coding in a state-of-the-art sinusoid-plus-noise based audio coder [4] and compared the coding results through listening tests with speech and audio signals. All signals used were mono and sampled at 44.1 kHz. The test excerpts were English female speech, castanets, a harpsichord solo, and pop music songs by Celine Dion and Eddie Rabbit. We studied the quality improvement of PLPC over standard LPC and the quality improvement of PLPC when combined with a flexible time segmentation and flexible model order allocation algorithm. Furthermore, we compared PLPC with the perceptual transform based on the masking threshold of the total input signal to the case where the transform was based on the masking threshold computed over the sinusoidal coded part only. Finally, we compared our method with the filterbank approach [3], after having removed the perceptual irrelevant sinusoids with the algorithm described in [5].

For the fixed segmentation cases, LPC or PLPC analysis/synthesis was applied on Hanning windowed segments of length 17.4 ms with 10th order LPC/PLPC. With flexible time segmentation and variable model order, the segments were Hanning windowed and allowed to vary from 5.8 ms to 23.2 ms in steps of 2.9 ms. The model orders were chosen from the set $\{4, 6, 8, 10\}$. In all cases, the target bit rate was set to 23 kbit/s, where 21 kbit/s was allocated to a sinusoidal coder and 2 kbit/s to the noise coder, including the segmentation overhead for the flexible time segmentation. We allocated 3 bits for each LPC-coefficient and 6 bits for the (log) LPC-gain. In situations with flexible model order, 2 bits per segment were allocated for model order information. In the case were we used the filter bank approach for coding the residual signal, the target bit rate was 24 kbit/s, of which 16 kbit/s was allocated to the sinusoidal coder and 8 kbit/s to the noise coder (see [4] for details). In our experiments, eight participants (the authors not included) listened to the original and five coded versions of the test excerpts. Every coded version had to be given a score from 1.0 (poor) to 5.0 (excellent).

Figure 3a shows the comparison between LPC and PLPC for fixed segmentation and fixed model order. We see that PLPC shows a better performance for all five fragments. A t-test with significance level $\alpha = 2.5\%$ reveals that for all fragments except the Eddie Rabbit fragment the differences in quality are statistically significant. Figure 3b shows the scores of PLPC with flexible segmentation and flexible order allocation and PLPC with fixed segmentation and fixed order allocation. For the castanets signal, which contains many transients, flexible segmentation leads to a statistically significant quality improvement, following the outcome of a t-test. For all other signals, the gain is not statistically significant. Figure 4a shows the results of flexible segmentation PLPC where the perceptual transformation is based on masking

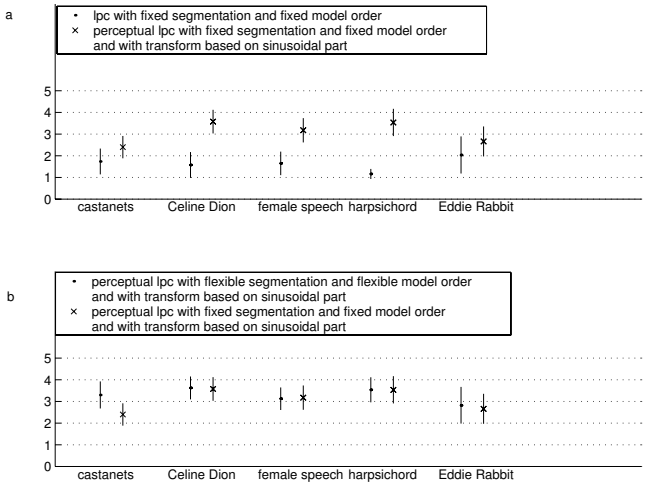

Fig. 3. *Experimental results averaged over all participants with accompanying standard deviation.*

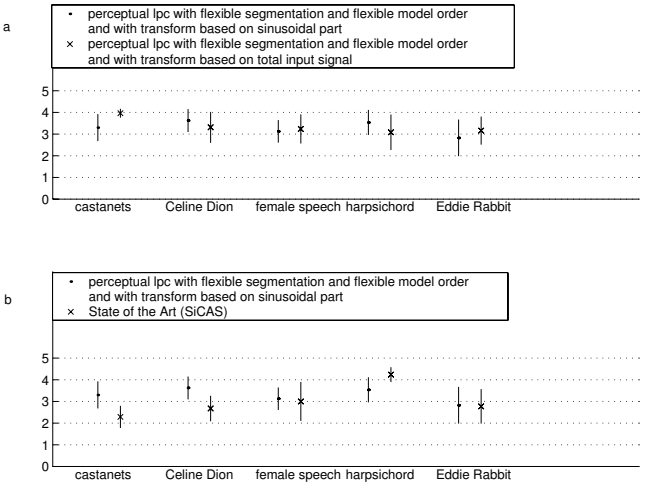

Fig. 4. *Experimental results averaged over all participants with accompanying standard deviation.*

curves computed over the total input signal compared to when the perceptual transform is based on masking curves computed over the sinusoidal coded part only. Both methods result in good scores with small differences, validating the assumption made in Section 3 that the transform can be based on the sinusoidal coded part only. In Figure 4b, PLPC with flexible time segmentation, variable model order and a transform based on the sinusoidal reconstruction only is compared to the filter-bank approach. A t-test of the data reveals that PLPC with flexible time segmentation and variable model order is significantly better compared to the state-of-the-art coder for the castanets and Celine Dion fragment. For the harpsichord fragment, both techniques perform very well with the filter bank approach having the best performance. For the female speech and Eddie Rabbit fragments both methods end up with an almost equal score.

6. CONCLUSIONS

We presented a new method to encode the residual after sinusoidal coding. This new method, called perceptual LPC (PLPC), is based on coding the residual signal after sinusoidal modelling in the perceptual domain by LPC. Where ordinary LPC minimizes a l_2-norm of the LPC modelling error to determine the optimal LPC-coefficients, this new method minimizes a perceptual norm of the modelling error. We showed that minimizing this perceptual norm is equivalent to minimizing a l_2-norm of the perceptually transformed modelling error, where the perceptual transform is based on the masking threshold. In contrast to other techniques aimed at coding the residual after sinusoidal coding, this method automatically models only the perceptually relevant noise-like components and does not waste bits on perceptually irrelevant sinusoidal components left after sinusoidal modelling. Listening tests showed that PLPC performs much better as a sinusoidal residual coding technique than ordinary LPC. We combined this method with a flexible time segmentation and variable model order allocation algorithm to be able to adapt the segmentation and model order to the underlying signal, leading to a significant gain in performance in comparison to a fixed time segmentation when applied to signals containing transients.

7. REFERENCES

[1] K. N. Hamdy, M. Ali, and A. H. Tewfik, "Low bit rate high quality audio coding with combined harmonic and wavelet representations," in *Proc. IEEE Int. Acoust., Speech and Signal Proc.*, 1996, pp. 1045–1048.

[2] R. Heusdens, R. Vafin, and W. B. Kleijn, "Sinusoidal modeling using psychoacoustic-adaptive matching pursuits," *IEEE Signal Processing Lett.*, vol. 9, no. 8, pp. 262–265, 2002.

[3] M. M. Goodwin, *Adaptive signal models*, Kluwer Academic Publishers, 1998.

[4] R. Heusdens et. al., "Sinusoidal coding of audio and speech (Si CAS)," *Submitted to Journal of the Audio Engineering Society*.

[5] N. H. van Schijndel, M. Gomez, and R. Heusdens, "Towards a better balance in sinusoidal plus stochastic representation," *Proc. IEEE workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 197–200.

[6] W. B. Kleijn and K. K. Paliwal, *Speech coding and synthesis*, chapter 12.

[7] H.Hermanski, B.A.Hanson, and H.Wakita, "Perceptually based linear predictive analysis of speech," in *Proc. IEEE Int. Acoust., Speech and Signal Proc.*, 1985, vol. 1, pp. 509–512.

[8] J. Makhoul, "Spectral linear prediction: properties and applications," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 3, pp. 283–296, June 1975.

[9] R. Heusdens and S van de Par, "Rate-distortion optimal sinusoidal modeling of audio and speech using psychoacoustical matching pursuits," in *Proc. IEEE Int. Acoust., Speech and Signal Proc.*, 2002, vol. 2, pp. 1809–1812.

[10] J. D. Markel and A. H. Gray jr., *Linear Prediction of Speech*, Springer-Verlag, 1976.

[11] P. Prandoni and M. Vetterli, "R/D optimal linear prediction," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 6, pp. 646–655, 2000.