A CANONICAL REPRESENTATION OF SPEECH

Mattias Nilsson, Barbara Resch, Moo-Young Kim, and W. Bastiaan Kleijn

KTH (Royal Institute of Technology)
School of Electrical Engineering
100 44 Stockholm, Sweden

ABSTRACT

It is well known that usage of an appropriate representation of the speech signal improves the performance of speech coders, recognizers, and synthesizers. In this paper we present a representation of speech that has the efficiency, in terms of being compact, similar to that of parametric modeling, but additionally has the completeness property of signal expansions. The resulting canonical representation of speech is suited for a wide range of speech processing applications and we demonstrate this through experiments related to coding and prosodic modification.

Index Terms— speech representation, perfect reconstruction, frame theory, energy concentration, best basis selection

1. INTRODUCTION

The representation of digitized speech by a set of parameters, each describing particular characteristics of the speech signal is of importance to all speech processing systems. The usage of an appropriate representation yields more efficient speech coding systems, improved quality of speech synthesis and (prosodic) modification systems, and increased performance of recognizers.

A general speech representation is specified by a model description, model parameters, and signal coefficients. If the representation consists of signal coefficients alone, such as pulse-code modulation (PCM) samples, it is commonly referred to as non-parametric, whereas a representation that is entirely model-based, such as the sinusoidal representation, is referred to as parametric.

It is beneficial to define a generic measure of goodness of a representation, independently of a particular application. A reasonable measure is compactness of the representation under the constraint of completeness. A compact (or sparse) representation implies a representation that has relatively few model parameters and has relatively few *significant* signal coefficients per unit time. This is consistent with the energy of the coefficient space being concentrated into a small subspace, which is advantageous for compression. By completeness of a representation we mean that the signal can be reconstructed perfectly given the representation. The concept of compactness and completeness of a representation was previously discussed in the context of speech in [1].

To arrive at a compact representation of a signal it is important to utilize all structure that the signal has. If we consider the speech signal in particular, we can identify two prominent regularities: short-term dependencies due to the resonances of the vocal tract, and long-term dependencies associated with the pitch (vibrations of the vocal folds).

Similarly to existing speech representations, e.g., [2, 3], the representation we propose utilizes both the short-term and long-term dependencies of the speech signal. However, a significant difference is that we actively search for a compact representation under the constraint of completeness (perfect reconstruction). The model that results in our compact and complete representation is a continuation of the ideas presented in [9].

The proposed system consists of four processing blocks: LP analysis, constant pitch warping, pitch-synchronous transform, and modulation transform. We select both the pitch-synchronous and modulation transforms to be frequency transforms. The combination of the pitch-synchronous and the modulation transforms form lapped frequency transforms [4], and similarly to all frequency transforms, they approximate the Karhunen-Loève transform (KLT) for stationary signal segments. The KLT maximizes the coding gain, which can be seen as a particular form of energy concentration in a statistical sense. (Note that this energy concentration in a statistical sense differs from the energy concentration for coefficients describing observed sequence, which we use elsewhere in this paper.)

If the pitch is constant, the pitch-synchronous and the modulation transforms can be applied directly on the speech signal to achieve a highly energy concentrated representation. It is desirable to describe the variances of the signal coefficients after the transformations in an efficient manner. Towards this goal we describe the spectral envelope by a parametric model. As is common in speech processing, we use the conventional AR model for this purpose. In practice, the pitch is varying over time, which implies that either the pitch-synchronous and modulation transforms have to adapt to the pitch, or the signal has to be made into a constant pitch signal.

In this work we warp the speech signal of varying pitch into a signal of constant pitch. The warping simplifies significantly the design of the pitch-synchronous and modulation transforms. The output of the warper is a description of a time-continuous pitch-track and a signal of constant pitch. Because of the constraint of completeness of the representation, the warped signal has to be oversampled, and thus, an increased efficiency of the AR-modeling and preservation of formant bandwidths and locations is obtained if the LP analysis is performed prior to the warper. This provides motivation for both the existence and the order of the processing blocks of the proposed system. An additional motivation for the proposed structure is that it facilitates the identification of voiced and unvoiced signal coefficients, which increases compactness, since a parametric description of the unvoiced signal coefficients suffices for good quality [5]. The decomposition in a voiced and an unvoiced component is also beneficial for prosodic modification (time- and pitch-scaling) of speech [6].

2. SPEECH ANALYSIS

As mentioned in the introduction, the proposed system consists of four stages: LP analysis, constant pitch warping, pitch-synchronous

transform, and modulation transform. The individual stages are described in detail in the following, starting with the linear-prediction analysis.

2.1. Linear-prediction analysis

To obtain a compact description of the distribution of energy along the frequency axis, we apply linear-prediction (LP) analysis to the signal. The LP analysis decomposes the signal into a set of time-varying parameters that specify an autoregressive (AR) model of the resonances and the spectral tilt, and a residual signal. Since the short-term statistics vary over time, the LP analysis is performed on short segments of the speech (typically of 20 ms duration). Let $s(n)$ denote sample n of a discrete speech signal. The LP analysis can then be expressed as $e(n) = s(n) + \sum_{m=1}^{M} a_{l,m}^{M} s(n-m)$, where $e(n)$ denotes the LP residual, and $a_{l,m}^{M}$ is the m'th linear prediction coefficient of prediction order M for a block l. The block index $l = \lceil n/L \rceil$, where $\lceil \cdot \rceil$ denotes the rounding upwards to the nearest integer, and L denotes the block length for the LP analysis.

2.2. Warping

For voiced speech the duration and shape of the pitch cycles generally change slowly. Thus, the long-term dependencies are strong (high redundancy), and can be used to obtain a compact representation. The first step in this process is warping the excitation $e(n)$ into a signal of constant pitch. That is, we separate the duration of pitch cycles (the pitch track) from the pitch-cycle shapes. The warping facilitates the following processing stages to yield a compact signal representation.

The warper is based on a continuous-time representation of the instantaneous pitch. In this work the warping function $t(\tau)$, relating the time domain t and the warped time domain τ, is modeled using cubic B-splines (with coefficients spaced proportional to the average pitch). The objective function of the warper is a waveform similarity measure between the signal and the signal delayed by one pitch period. We seek the B-spline coefficients of the warping function $t(\tau)$ that minimizes the objective function. For a detailed description of the estimation of the instantaneous pitch and the warping function we refer to [7].

Because of the one-to-one mapping between the warped and original time-domains, we can derive the inverse mapping $\tau(t)$ from the estimate of $t(\tau)$. $\tau(t)$ is needed for the reconstruction of the original residual from the warped residual, which is discussed in Section 3.

The discrete warped signal sample $e_{warped}(\nu)$, where ν is the sample index, can be expressed as

$$e_{warped}(\nu) = \Theta e(\nu) = \langle e, \theta_\nu \rangle, \ \forall \nu \in \mathbb{Z}, \tag{1}$$

where Θ is a frame operator [8] applied to the residual signal e of infinite dimension, and θ_ν is the ν'th frame function. The warping is effectively an irregular sampling; under certain conditions the set of displaced sinc functions form a frame [8]. The frame functions θ_ν are defined as $\theta_\nu(n) = \mathrm{sinc}(n - t(\nu))$, where $\mathrm{sinc}(x) = \sin(\pi x)/(\pi x)$. The warped residual at the discrete warped sample index ν becomes $e_{warped}(\nu) = \sum_{n \in \mathbb{Z}} e(n)\theta_\nu(n)$. Thus, the discrete warped residual e_{warped} is formed by an irregular (over-) sampling of the continuous-time residual $e(t)$.

2.3. Pitch-synchronous transform

The outputs of the warper are a residual signal of constant pitch and a description of the corresponding continuous-time pitch-track. By utilizing the high redundancy between consecutive pitch cycles it is possible to get a compact representation. This, for instance, can be accomplished by a two-stage procedure with a pitch-synchronous transform followed by a modulation transform (cf. the two-stage sparseness approach of [9]).

In this work, we select a modulated lapped transform (MLT) [4] as the pitch-synchronous transform. The MLT has the advantage of facilitating a critically sampled uniform filter bank with coefficients that are localized in time and frequency.

Let Φ denote the frame operator. The MLT coefficients can then be expressed as

$$f(k,l) = \Phi e_{warped}(k,l) = \\ \langle e_{warped}, \phi_{kl} \rangle, \ \forall k \in \mathbb{Z}, \forall l \in \{0, ..., P_0 - 1\}, \tag{2}$$

where k and l are time and frequency indices, respectively. In our implementation, the frame functions ϕ_{kl}'s are constructed from the combination of the square-root of a Hann window (to satisfy the power complementarity constraint needed for the perfect reconstruction) and DCT-IV functions. That is,

$$\phi_{kl}(\nu) = w_\nu \sqrt{\frac{2}{P_0}} \cos\left(\frac{(2l+1)(2\nu - (2k+1)P_0 + 1)\pi}{4P_0} \right), \tag{3}$$

where P_0 denotes the normalized pitch, and where the window w_ν has a non-zero support only for $\nu \in [kP_0, ..., (k+2)P_0 - 1]$.

2.4. Modulation transform

If we consider a speech signal belonging to a steady voiced sound, the warped residual consists of a sequence of similarly shaped pitch cycles of equal duration. The output from the pitch-synchronous transform of such a signal is a sequence of MLT coefficients that change slowly over time. Thus, applying a modulation transform on this sequence renders signal coefficients that are compact, i.e., the energy of the modulation transform coefficients is concentrated into the lowest modulation bands. This is the main motivation for a modulation transform.

An additional motivation is that a modulation transform allows for identification of voiced and unvoiced signal coefficients, beneficial for both coding and modification of speech. For the voiced-unvoiced decomposition we assign the low modulation bands to our voiced speech category. The coefficients of the low modulation bands represent the constant and slowly evolving components of the pitch-synchronous coefficients over time (block length of the modulation transform). In this work we assign the lowest 20% (minimum three bands) of the modulation bands to belong to the voiced category.

Our implementation of the modulation transform is a DCT-II with a rectangular window of adaptive length. The combination of rectangular windows and DCT-II facilitates the implementation of the modulation transform as a critically sampled filter bank. A desired property of the resulting filter bank, from an energy concentration point of view, is that a sequence of constant coefficients from the pitch-synchronous transform renders only the DC coefficient nonzero in the modulation domain. The selection of the window lengths is based on an energy concentration criterion of the modulation coefficients, assigning short windows (high temporal resolution) to rapidly changing regions and long windows to steady regions. Thus, our adaptive modulation transform is a best basis selection using the local cosine basis functions of the DCT. Given the window lengths from the best basis selection the modulation transform can be expressed as a set of frames, where the modulation transform

coefficients $g(p,q,l)$ are formed by the inner products between the pitch-synchronous coefficients and the modulation frame-functions, i.e.,

$$g(p,q,l) = \Psi f(p,q,l) = \langle f, \psi_{pql} \rangle, \\ \forall p \in \mathbb{Z}, \forall q \in \{0,...,Q_p - 1\}, \forall l \in \{0,...,P_0 - 1\}, \quad (4)$$

where Ψ denotes the modulation frame operator and p, q, and l denote time block, modulation frequency, and frequency bands, respectively.

The frame function ψ_{pql} is constructed by a set of windowed DCT-II functions and defined as

$$\psi_{pql} = v_{Q_p} c(q) \sqrt{\frac{2}{Q_p}} \cos\left(\frac{(2(k - \sum_{j=1}^{p-1} Q_j) + 1) q \pi}{2 Q_p} \right), \quad (5)$$

where the modulation bands $q \in \{0,...,Q_p - 1\}$, and where v_{Q_p} denotes a rectangular window of effective length Q_p starting at position $k - \sum_{j=1}^{p-1} Q_j$, and $c(0) = 1/\sqrt{2}$ and $c(q) = 1$ for $q \neq 0$.

Similarly to the best basis algorithms presented in [10], we only consider a computationally efficient method for the selection of the set of $\{Q_p\}_{p \in \mathbb{Z}}$. The initial length of the modulation window for block p is set to one, i.e., $Q_p = 1$ and then extended as long as there is an increase in the energy concentration of the corresponding modulation coefficients. Many energy concentration cost functions can and have been used, e.g., coding gain based [11] or entropy based [10], and the particular choice depends on the application at hand. In our application we found the use of the entropy based energy concentration criterion of [10] inappropriate since the modulation coefficients of the highest frequency channels contain essentially no energy (due to the oversampling of the warper), and therefore, have a too large impact on the criterion. Thus, to measure the energy concentration of the modulation coefficients of block p with window size Q_p, we use the following function

$$C_E(p, Q_p) = - \sum_{q=0}^{Q_p - 1} \sqrt{\sum_{l=0}^{P_0 - 1} g(p,q,l)^2}, \quad (6)$$

and we extend the window size Q_p to $Q_p + 1$ if

$$C_E(p, Q_p + 1) \geq C_E(p, Q_p) + C_E(p+1, 1) - \lambda. \quad (7)$$

The λ in (7) is a bias-term favoring long windows which is advantageous for silence and purely unvoiced sounds to have a larger portion of the modulation coefficients assigned as unvoiced (since we assign to the voiced speech category the coefficients of at least the three lowest modulation bands). For wideband speech represented in a raw 16 bit format (amplitudes from -32767 to 32768) we found $\lambda = 1000$ to be an acceptable choice.

Fig. 1 shows an example of the behavior of the transforms and voiced/unvoiced decomposition of a voiced onset. In Fig. 1 note how the segment labeled B isolates the transient-like onset, and that the segment labeled C captures the steady voiced region.

3. SPEECH SYNTHESIS

In the speech analysis operation we separated the coefficients of the modulation transform into the voiced and unvoiced categories. The synthesis operation, from the modulation coefficients to the linear-prediction residual, consists of a series of expansions using the dual frame of each forward frame expansion. That is, the linear-prediction residual from, e.g., the *voiced* modulation coefficients, g_{voiced}, can be expressed as

$$e_{voiced}(n) = \Theta^\sharp \Phi^\sharp \Psi^\sharp g_{voiced}(n), \ \forall n \in \mathbb{Z}. \quad (8)$$

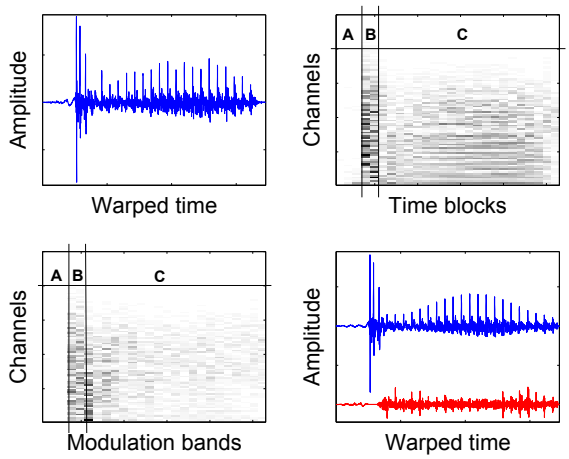

Fig. 1. Top left: Warped residual at a voiced onset. Top right: Intensity plot of the corresponding MLT coefficients from the pitch-synchronous transform. The A, B, and C labels the segments obtained from the best basis selection (A is incompletely displayed as B and C are of main interest.). Bottom left: Intensity plot of the coefficients from the modulation transform (note that the low modulation bands are located to the left in each of the segments A,B, and C). Bottom right: reconstructed voiced and unvoiced warped residuals.

In more detail, the *voiced* MLT coefficients of the pitch-synchronous transform are obtained from the corresponding modulation coefficients by the expansion $f_{voiced}(k,l) = \sum_{p \in \mathbb{Z}} \sum_{q=0}^{Q_p - 1} g(p,q,l) \psi_{pql}(k)$. Similarly, the warped residual belonging to the voiced MLT coefficients becomes $e_{warped,voiced}(\nu) = \sum_{k \in \mathbb{Z}} \sum_{l=0}^{P_0 - 1} f_{voiced}(k,l) \phi_{kl}(\nu)$. Note that the pitch-synchronous and modulation transforms are both *tight* frames (or orthogonal bases) and that, thus, their inverses are trivial. The warped residual is an irregular oversampling of the original residual (cf. Section 2). The inverse frame operator is the pseudo inverse of the analysis frame. Thus, the voiced LP residual is $e_{voiced}(n) = \sum_{\nu \in \mathbb{Z}} e_{warped,voiced}(\nu) \theta_n^\sharp(\nu)$, where $\theta_n^\sharp$ is the n'th row vector of the pseudo-inverse $\Theta^\sharp = (\Theta^H \Theta)^{-1} \Theta^H$.

Finally, applying the LP synthesis filter to the residual yields the voiced speech signal. The reconstruction of the unvoiced speech signal is analogous.

4. EXPERIMENTS AND RESULTS

Both coders and systems for time- and pitch-scaling of speech can be made based on our proposed speech representation. For instance, a variable-rate speech coder can be developed by applying entropy-constrained quantizers to the parameters and coefficients followed by an arithmetic coder. A prosodic modification system for time- and pitch-scaling of speech can be obtained by interpolating the pitch-synchronous coefficients and changing the warping function $t(\tau)$ correspondingly.

For both speech coding and prosodic modification, voiced-unvoiced decomposition is of great importance. In coding we can achieve large coding gains by replacing the unvoiced waveform descriptors with Gaussian noise (with properly matched gain and color) [5]. Prosodic modification systems also benefit from the decomposition since artificial periodicity introduced when stretching unvoiced

sounds in time can be combated by e.g. randomizing the phase of the unvoiced signal components. Thus, in this section we demonstrate the performance of the voiced-unvoiced decomposition of our system. However, we start with providing some implementation specific details.

4.1. Implementation specific details

For all the experiments we used speech from the TIMIT database sampled at 16 kHz. The LP analysis is performed every 20 ms using a prediction order of 18 with a bandwidth expansion factor of 0.997. The normalized pitch lag P_0 is set to 256. To reduce computational complexity we approximate the pseudo-inverse by an irregular (down-) sampling when going from the warped to the regular time domain. We obtain an signal to noise ratio of approximately 60-70 dB between the original and synthesized speech using bandlimited sinc-interpolation (Hann window of 100 samples support).

4.2. Voiced fricatives

The decomposition into voiced and unvoiced components is done continuously over the time and frequency space. This facilitates voiced-unvoiced separation even for voiced fricatives. Voiced fricatives are fricatives articulated with oscillating vocal folds. Fig. 2 shows the original (top), voiced (middle), and unvoiced (bottom) power spectra of a /z/-sound. The strong harmonic character of the decomposed voiced signal spectrum is clearly visible in the voiced power spectrum and it is interesting to note that the unvoiced signal component dominates the original spectrum already above 1 kHz.

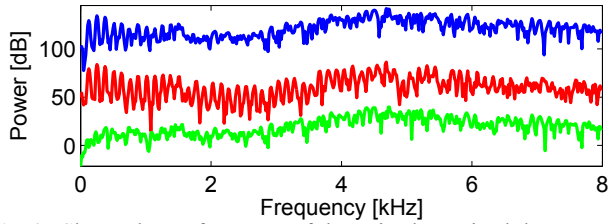

Fig. 2. Shows the performance of the voiced-unvoiced decomposition of a voiced fricative /z/-sound. From top to bottom we have the original, the voiced, and the unvoiced power spectra, respectively.

4.3. Randomized unvoiced components

As mentioned above, proper voiced-unvoiced decomposition is beneficial for both speech coding and prosodic modification. Therefore, to test the performance of the decomposition we randomize the sign of the unvoiced pitch-synchronous coefficients. To preserve the distribution of energy within a pitch cycle of the unvoiced residual over time, we ensure that the randomized and original unvoiced residuals have similar smoothed amplitude-envelope (absolute of Hilbert transformed signal, convolved with Hann window of 20 samples support, and raised to power 1.3). The modulation is only applied to the frequencies above 0.5 kHz. For robustness to errors in the warping-function we assign all components below 1 kHz to be voiced if the first reflection coefficient of the corresponding time-block is below -0.7.

To evaluate the perceptual quality of the system when randomizing the unvoiced signal, as described above, we performed the Comparison Category Rating (CCR) [12]. As anchors we used the modulated noise reference units (MNRU) [13] at SNR levels 25 and 30 dB (labeled Q.25 respectively Q.30 in the following). From each of the eight dialect regions of the TIMIT database we randomly selected one male and one female speech sentence for the test. The eight subjects that participated in the listening test were asked to grade the quality of the processed speech (i.e., ours denoted by CRoS, and the anchors Q.25 and Q.30) compared to the original on an integer scale from -3 to 3 corresponding to *much worse*, *worse*, *slightly worse*, *equal*, *slightly better*, *better*, and *much better* [12]. Each pair of sentences were played twice and in random order to increase the statistical significance of the test. The results from the listening test are displayed in Table 1 and show that our system is rated between equal and slightly worse compared to the original.

CRoS	Q.25	Q.30
-0.45 +/- 0.14	-1.64 +/- 0.11	-0.94 +/- 0.12

Table 1. Mean scores together with 95% confidence intervals of the CCR listening test when comparing CRoS, Q.25, and Q.30 to the original speech. Eight subjects participated in the test.

5. CONCLUDING REMARKS

In this paper we describe a method that uses the redundancies in the speech signal to derive a compact and complete representation of digitized speech. The representation is mathematically tractable, guarantees perfect reconstruction, and is suitable for a wide range of speech processing applications. Through experiments we have shown the efficiency of the voiced and unvoiced separation of our systems. The method forms a strong foundation for speech coders and systems for time- and pitch-scaling.

6. REFERENCES

[1] W. B. Kleijn and D. Talkin, "Compact speech representations for speech synthesis," in *Proc. IEEE Workshop on Speech Synthesis*, 2002, pp. 35–38.

[2] J. Laroche, Y. Stylianou, and E. Moulines, "HNM: A simple, efficient harmonic + noise model for speech," in *IEEE Workshop on Applications of Sign. Proc. to Audio and Acoust.*, 1993, pp. 169–172.

[3] G. Evangelista, "Pitch-synchronous wavelet representation of speech and music signals," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3313–3330, 1993.

[4] H. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 38, no. 6, pp. 969–978, 1990.

[5] G. Kubin, B. S. Atal, and W. B. Kleijn, "Performance of noise excitation for unvoiced speech," in *IEEE Workshop on Speech Coding for Telecommunications*, 1993, pp. 35–36.

[6] Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic + noise model," in *Proc. Eurospeech*, 1995, pp. 451–454.

[7] B. Resch, A. Ekman, M. Nilsson, and W. B. Kleijn, "Estimation of the instantaneous pitch of speech," *Accepted for publication in IEEE Trans. on Speech, Audio, and Language Processing*, 2006.

[8] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1998.

[9] W. B. Kleijn, "A frame interpretation of sinusoidal coding and waveform interpolation," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, 2000, vol. 3, pp. 1475–1478.

[10] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Information Theory*, vol. 38, no. 2, pp. 713–718, 1992.

[11] N. S. Jayant and P. Noll, *Digital coding of waveforms*, Prentice Hall, New Jersey, USA, 1984.

[12] "Methods for subjective determination of transmisson quality," ITU-T Recommendation P.800, August 1996.

[13] "Modulated noise reference unit (MNRU)," ITU-T Recommendation P.810, February 1996.