Excitation Signal Generation in a Speech Reconstruction System

Patrick Hannon, Bernd Iser, Mohamed Krini
SVOX Deutschland GmbH, 89077 Ulm, Deutschland, Email: patrick.hannon@svox.com

Introduction

Speech reconstruction attempts to recreate regions of a speech signal that are corrupted due to a low signal-to-noise ratio (SNR) and are consequently also suppressed by the noise suppression module of a speech enhancement system. In contrast to the work presented in [1], a method to reconstruct a speech signal in the frequency domain is desired to avoid the complexities of time-domain pitch-synchronous-overlap-add (TD-PSOLA) synthesis [2]. The first proposed step is to train models of clean voiced speech and then apply the models to quantize corrupted excitation vectors that are extracted from voiced frames of noisy input speech. The quantized vectors are then used in the speech reconstruction and this paper describes investigations into the performance and viability of this proposal by comparing combinations of quantization methods on clean input speech.

Speech Excitation Model

The Short Time Fourier Transform (STFT) of the real-valued discrete-time input signal $x(d)$ is defined as

$$X_w\left(e^{j\Omega_\mu},n\right) = \sum_{m=0}^{M-1} x(m+nR)w_{\mathrm{a}}(m)e^{-j\Omega_\mu m} \qquad (1)$$

for $\mu = 0,\ldots,\tilde{M}-1$, where n is the frame index, the frameshift is $R = 160$, and w_a is an analysis window function of length $M = \tilde{M} = 512$.

The source-filter model of speech production [3] proposes that a glottal excitation signal is filtered by the vocal tract. The vocal tract can be represented by the magnitude spectral envelope, $|H(e^{j\Omega_\mu},n)|$, which is removed from $X_w\left(e^{j\Omega_\mu},n\right)$ by

$$Q\left(e^{j\Omega_\mu},n\right) = \frac{X_w\left(e^{j\Omega_\mu},n\right)}{|H(e^{j\Omega_\mu},n)|} \qquad (2)$$

to obtain an estimate of the complex-valued excitation signal, $Q(e^{j\Omega_\mu},n)$.

Consequently, the excitation magnitude spectrum is defined as $|Q(e^{j\Omega_\mu},n)|$ and the excitation phase spectrum, $\theta_\mu(n)$, is calulated by $\arg(Q(e^{j\Omega_\mu},n))$ so that

$$X_w\left(e^{j\Omega_\mu},n\right) = \left|H\left(e^{j\Omega_\mu},n\right)\right| \left|Q\left(e^{j\Omega_\mu},n\right)\right| e^{j\theta_\mu(n)}. \qquad (3)$$

Excitation Magnitude & Phase Estimates

Before speech reconstruction can take place, estimates of $Q(e^{j\Omega_\mu},n)$ and $\theta_\mu(n)$ must be calculated. The excitation magnitude estimate, $|\hat{Q}(e^{j\Omega_\mu},n)|$, is derived directly from $Q(e^{j\Omega_\mu},n)$. However, instead of estimating the excitation phase directly, the dynamic phase shift of each STFT frequency bin from one frame to the next is extracted by

$$\Delta\theta_\mu(n) = \theta_\mu(n) - \theta_\mu(n-1). \qquad (4)$$

An estimate, $\Delta\hat{\theta}_\mu(n)$, of this phase shift spectrum is then added to the phase estimate of the previous frame as defined by

$$\hat{\theta}_\mu(n) = \hat{\theta}_\mu(n-1) + \Delta\hat{\theta}_\mu(n) \qquad (5)$$

to calculate the phase estimate of the current frame. By modeling the frameshift-based phase differences of each frequency bin, the temporal phase incoherencies that would result from static phase estimation can be reduced in overlapping frame-based processing.

Substituting $|\hat{Q}(e^{j\Omega_\mu},n)|$ and $\hat{\theta}_\mu(n)$ into Eq. 3 produces the synthetic voiced speech signal, $\hat{X}_w(e^{j\Omega_\mu},n)$. Estimates for the excitation magnitude and phase spectra are calculated using vector quantization codebooks and a short description of this method is presented in the next section.

Vector Quantization Codebook Training

Frames of voiced speech with pitch, $f_0(n)$, are sorted into pitch intervals of size Δf_0. Only frames whose pitch is within a limited range of interest, 80 Hz $\leq f_0(n) <$ 260 Hz, are considered. Inside each pitch interval, K clusters are trained using a clustering algorithm based on the Linde-Buzo-Gray (LBG) algorithm from [4] where the number of clusters is increased by 1 in each iteration (from 2 to K), each time splitting the cluster with the largest sum of distances for the next iteration. After $K-1$ iterations, the k^{th} codebook entry is defined as the cluster member with the smallest distance to the k^{th} cluster center.

To analyze the effects of pitch interval size, 3 codebook versions are trained for $\Delta f_0 \in \{5, 10, 20\}$Hz. Further investigations involve the effects from the number of clusters, K, and 3 codebook versions are trained for class sizes of $K \in \{5, 8, 16\}$ and each size of Δf_0.

A comparison between independent and synchronized training methods for the excitation magnitude and phase shift codebooks is also presented. For the independent method, optimized codebooks for $|Q(e^{j\Omega_\mu},n)|$ and $\Delta\theta_\mu(n)$ vectors are trained separately. Conversely, the synchronized method ensures that the k^{th} codebook entry for both the magnitude and phase shift codebooks originate from the same frame of training data by using supervectors of $\left|Q\left(e^{j\Omega_\mu},n\right)\right|$ and $\Delta\theta_\mu(n)$ as training vectors.

Before training codebooks involving phase shift, the effects of phase wrapping need to be compensated. This compensation is performed before computing distances, while retaining the original phase shift vectors to be used as codebook entries. The same precaution must also be met during speech reconstruction, which is described in the next section.

Codebook Based Speech Reconstruction

During speech reconstruction, the excitation magnitude and phase shift spectra of an input signal frame are first extracted. Then after assigning the frame to the correct Δf_0 interval, the LBG-based quantization is performed by finding the k^{th} entry with the smallest quantization error. Quantization error is measured up to $M_{\max}$, which represents the maximum frequency bin of interest during voiced speech, corresponding here to an approximate frequency of 3300 Hz. Excitation magnitude quantization error is measured using squared Euclidian distance while the quantization error of excitation phase shift is measured using the delta phase distance (DPD) presented in Eq. 7. The distance measures are used for both the codebook training and quantization steps.

For the independent codebooks, these error values are calulated separately for the excitation magnitude and phase shift codebooks. In the case of the synchronized codebooks, the error is calculated by the sum of the separate excitation magnitude and phase shift error values.

Results

Errors in excitation magnitude quantization, are shown in Table 1 and are presented as average log spectral distances (LSD) defined by

$$LSD = \frac{10}{N}\sum_{n=0}^{N-1}\sqrt{\frac{1}{M_{\max}}\sum_{\mu=0}^{M_{\max}-1}\log_{10}^2\frac{(|Q(e^{j\Omega_\mu},n)|)^2}{(|\hat{Q}(e^{j\Omega_\mu},n)|)^2}}. \quad (6)$$

As expected, the error generally increases with increasing Δf_0 size and decreases for a higher number of K classes, and this trend is found in both the independent and the synchronized codebooks.

Table 1: Excitation magnitude error - LSD values in dB

Variations	Δf_0 in Hz		
	5	10	20
$K = 5$ ind.	3.2	3.3	3.5
$K = 8$ ind.	3.2	3.1	3.3
$K = 16$ ind.	3.1	3.1	3.1
$K = 5$ sync.	3.3	3.6	3.9
$K = 8$ sync.	3.4	3.5	3.6
$K = 16$ sync.	3.2	3.3	3.3

Average phase shift quantization errors, seen in Table 2 are measured using a modified Euclidean distance defined as the delta phase distance (DPD) calculated by

$$DPD = \frac{10}{N}\sum_{n=0}^{N-1}\frac{1}{M_{\max}}\sum_{\mu=0}^{M_{\max}-1}\sin^2\left(\frac{\Delta\theta_\mu(n)-\Delta\hat{\theta}_\mu(n)}{2}\right). \quad (7)$$

This formula produces normalized error values so that the largest phase shift quantization distances, i.e., multiples of $-\pi$ and π, are normalized to 1. The values follow the same pattern as the LSD values above in relation to Δf_0 size and number of K classes. Although the variation of the error values is small, it can be seen that for the synchronous training method, error is reduced compared to the independent training version.

Conclusions

LBG-based codebooks perform well for excitation magnitude in terms of the LSD measure. Increased codebook resolution delivers improved performance and error is slightly higher for the synchronized codebooks than for indepedent codebooks. This is the expected tradeoff for considering the phase shift vectors simultaneously.

Table 2: Phase shift error - DPD values (normalized to 1)

Variante	Δf_0 in Hz		
	5	10	20
$K = 5$ ind.	0.39	0.40	0.41
$K = 8$ ind.	0.39	0.39	0.41
$K = 16$ ind.	0.39	0.39	0.40
$K = 5$ sync.	0.37	0.37	0.39
$K = 8$ sync.	0.37	0.37	0.37
$K = 16$ sync.	0.37	0.37	0.37

Excitation phase shift quantization error in the synchronous training method is reduced compared to the independent training version. Thus, introducing a dependency on the excitation magnitude for selection of the appropriate phase shift codebook entry improves performance.

In informal listening comparisons, files constructed using a Δf_0 size of 20 Hz sound very unnatural. This can be attributed to the large transitions between frames whose $f_0(n)$ are more than 20 Hz different from each other. Similarly, the comparatively poor resolution introduces an undesired uniformity for frames where the differences in f_0 are smaller than 20 Hz.

Outlook

The results indicate a need to develop a revised method of codebook training. Alternative pre-clustering methods, which more thoroughly exploit dynamic features, are expected to improve quantization performance. New experiments are already underway and have begun to deliver promising results. Additionally, the DPD distance measure must be explored and possibly improved.

Furthermore, while the $\Delta\theta_\mu(n)$ vectors concentrate on inter-frame coherency, it is also considered relevant to preserve the intra-frame coherency of the instantaneous phase differences over frequency. Such a feature vector could be calulated by

$$\Delta\phi_\mu(n) = \theta_\mu(n) - \theta_{\mu-1}(n). \quad (8)$$

This feature vector could then also be integrated into the training algorithms to preserve both inter-frame and intra-frame coherency.

References

[1] Krini, M. & Schmidt, G.: "Model-Based Speech Enhancement", in E. Hänsler, G. Schmidt (eds.), *Speech and Audio Processing in Adverse Environments*, Berlin, Germany: Springer, pp. 89-134, 2008

[2] Moulines, E. & Charpentier, F.: "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5-6, pp. 453-467, 1990

[3] Vary, P. & Martin, R.: *Digital Speech Transmission: Enhancement, Coding And Error Concealment*, John Wiley & Sons, pp. 10-25, 2006

[4] Linde, Y., Buzo, A. & Gray, R.: "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Commun.*, vol. 28, no. 1, pp. 84-94, 1980