Time-frequency voiced and unvoiced excitation models for harmonic speech systems

Miguel Arjona Ramírez

Abstract

Time-frequency voiced and unvoiced models are proposed for the excitation of a harmonic autoregressive wideband speech analysis-synthesis system. The time-frequency voiced excitation (TFVEX) model has low time resolution defined by the concentration of the excitation signal distribution in the modulation domain while the time-frequency unvoiced excitation (TFUNEX) model has cycle time discrimination with lower amplitude resolution and while the frequency resolution for both models is an octave. The speech reconstructed by the compound TFUVEX unvoiced-voiced model is rated above the speech degraded by a modulated noise reference unit (MNRU) at 25 dB in listening tests while yielding a parametric compression of over ten times.

Index Terms

speech analysis, speech coding, sparse representations, modulation transform, time-frequency analysis, voiced-unvoiced decision

This work is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) under Grant no. 307633/2011-0 and by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) under Grant no. 2009/18242-5.

Time-frequency voiced and unvoiced excitation models for harmonic speech systems

Time-frequency voiced and unvoiced excitation models for harmonic speech systems

I. Introduction

HARMONIC speech representations have been used for coding at medium to low bit rates [1] even when they fail to achieve perfect reconstruction. However, it is desirable to have a nearly perfect reconstruction (NPR) front-end representation since it is not bounded in performance at high rates and it will be more useful if amenable to be fit by simpler models for operation at lower bit rates that should be controlled by a manageable and meaningful set of parameters.

The classification of speech segments into voiced and unvoiced classes is important for speech modification and speech coding since they are processed differently. Besides, sparse speech representations are useful for source separation [2] and, if conceptually designed, they may be the basis for pattern playback in signal processing education [3].

Usually, a model is fit to the voiced harmonic amplitudes [1], [4] for ease of manipulation and often as an intermediate stage in coding. Somehow unexpectedly, an unvoiced model is important in making a harmonic system deliver natural-sounding speech [1], [5] or, for that matter, any synthesized audio signal [6], in particular music [7], [8] since any instrumental performance requires some random fluctuations in order to sound natural.

In this work voiced and unvoiced models are proposed within a common framework for an NPR front-end representation [9] so that sparse speech representations may be achieved.Unlike usual harmonic representations which apply hard decision for voiced and unvoiced speech classification in the time and/or in the frequency domain, the voiced-unvoiced decision in this analysis-synthesis system (ASyS) is equivalent to a soft decision because the separation is implemented in the modulation domain.

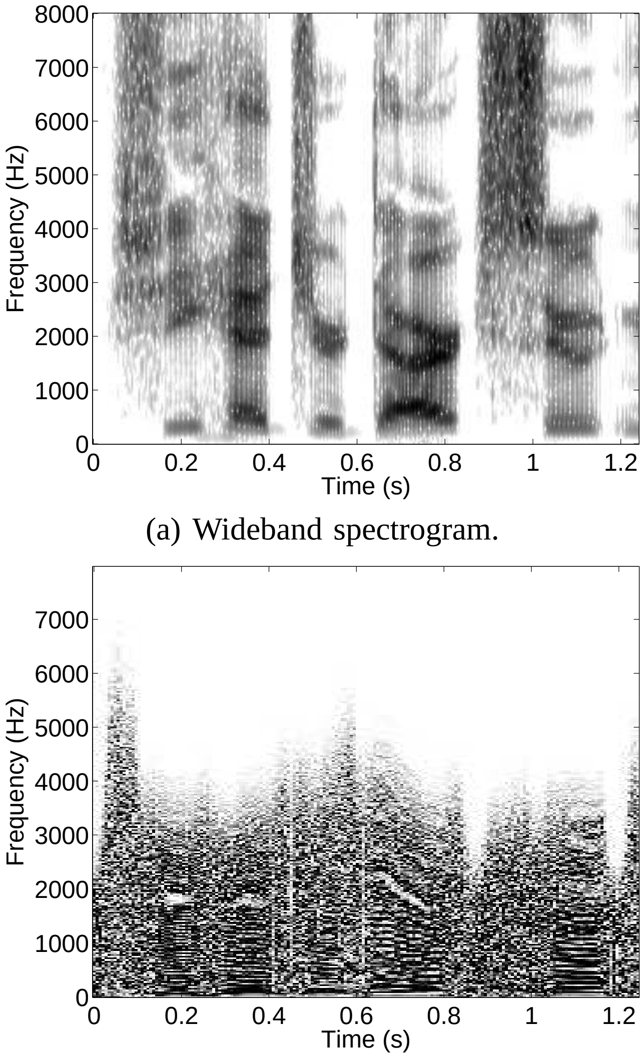
A brief description of ASyS and the criteria for classifying and segmenting speech excitation into voiced part and unvoiced part are presented in Section II. The voiced model (TFVEX) is developed in Section III upon specific time-frequency features of voiced excitation, which are to be contrasted with those of the unvoiced model (TFUNEX) that follows in Section IV. Then, the compound voiced-unvoiced model (TFUVEX) is presented in Section V with the introduction of the spectral weighting in the pitch-synchronous domain. Finally, the three excitation models are assessed in Section VI and remarks are drawn in conclusion.

II. Speech excitation classification and separation

A lower dynamic range in the time-frequency domain can be achieved by prewhitening the speech signal by means of linear prediction [10]. At a sampling rate of 16 kHz, 18th-order AR modeling has been found adequate with von Hann window length of 20 ms at a frame rate of 400 Hz. The prediction residual signal $r(n)$ is then time-warped to $r_w(\nu)$, for $\nu \in \mathbb{Z}$, by means of bandlimited sinc functions in order to hold the pitch period length constant at P_0 samples [9] as depicted in Fig. 2 whereas the original pitch track $p(n)$ is separated. Consequently, at this point the signal power is more evenly distributed in both the time and the frequency domains as illustrated in Fig. 1, where the time-frequency distributions of the power in the speech signal and in the residual signal are represented by a spectrogram and by the intensity of a pitch-synchronous transform (PST), respectively.

The time-warped residual undergoes a pitch-synchronous transform, which is a modulated lapped transform (MLT) [11] that produces the harmonic tracks

$$c_l(k) = \sum_{\nu=kP_0}^{(k+2)P_0-1} r_w(\nu)\phi_l\left(\mu_k(\nu)\right) \tag{1}$$



(a) Wideband spectrogram.

(b) Pitch-synchronous transform intensity plot.

Fig. 1. Wideband spectrogram of the speech signal and PST intensity plot of the prediction residual for the phrase “She had your dark suit...”, uttered by a male speaker.

for pitch cycles $k \in \mathbb{Z}$, harmonic indices $l \in \{0, 1, \dots P_0 - 1\}$ and local warped time $\mu_k(\nu) = \nu + [4 - (k \mod 4)]\, 2P_0$, where the basis functions are

$$\phi_l(\mu_k) = \sqrt{\frac{2}{P_0}} \cos\left[\frac{2\pi\left(l + \frac{1}{2}\right)\left(\mu_k - \frac{P_0}{2} + \frac{1}{2}\right)}{2P_0}\right] w_k\left(\mu_k\right). \tag{2}$$

The PST is evaluated at each pitch cycle over a two-cycle long window $w_k(\mu_k)$, which is a modified square-root von Hann window

$$w_k\left(\mu_k(\nu)\right) = \sqrt{\frac{1}{2} - \frac{1}{2}\cos\left[\frac{2\pi\left(\mu_k(\nu) + \frac{1}{2}\right)}{2P_0}\right]} \tag{3}$$

for the kth pitch cycle. In fact, any smooth window could be used as long as its squared sum is unity so that it satisfies the perfect reconstruction condition.

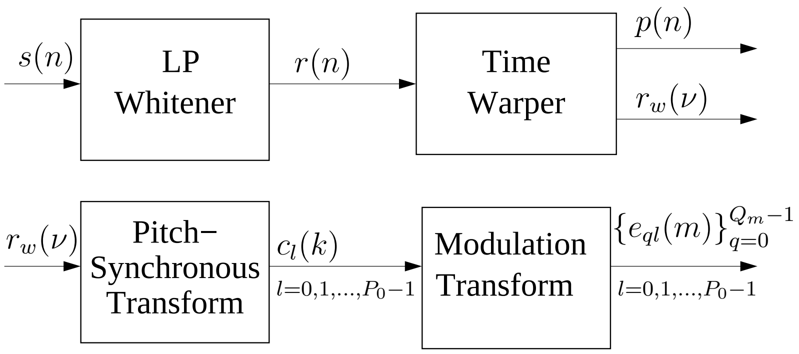


Fig. 2. Block diagram of the analysis process in ASyS.

The modulation transforms (MTs) of the resulting PST time tracks, for $l \in \{0, 1, \dots P_0 - 1\}$, are obtained in modulation segment m as

$$e_{ql}(m) = \sum_{k=k_{0m}}^{k_{0m}+Q_m-1} c_l(k)\psi_{mq}\left(\chi_m(k)\right) \tag{4}$$

where Q_m is the segment length with initial cycle index k_{0m}, local cycle index $\chi_m(k) = k - \sum_{i=0}^{m-1} Q_i$ and type-II DCT [12], [13] basis functions

$$\psi_{mq}(\chi_m) = \sqrt{\frac{2}{Q_m}}\gamma_q \cos\left(\frac{q\pi}{Q_m}\chi_m + \frac{q\pi}{2Q_m}\right) w_m(\chi_m) \tag{5}$$

for modulation frequencies (MFs) $q \in \{0, 1, \dots, Q_m - 1\}$, where coefficient γ_q is defined as

$$\gamma_q = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } q = 0 \mod Q_m \\ 1 & \text{if } q \neq 0 \mod Q_m \end{cases}$$

and $w_m\left(\chi_m(k)\right)$ represents the rectangular window for the mth modulation segment, which is supported within the interval $\{k_{0m}, k_{0m} + 1, \dots, k_{0m} + Q_m - 1\}$.

It should be noted that the analysis process shown in Fig. 2 can be inverted perfectly for obtaining the excitation signal back. The MT and the PST can be inverted perfectly within the numerical precision of the system and the LP-whitener plus time-warper cascade couple can be inverted with segmental SNR of around 50 dB for obtaining the reconstructed speech signal.

The time-frequency distribution concentration (TFDC) measure used for the modulation transform is the unnormalized modified Zakai's entropy

$$C_E(m, Q_m) = -\sum_{q=0}^{Q_m-1} \sqrt{\sum_{l=0}^{P_0-1} e_{ql}^2(m)} \tag{6}$$

where the square root replaces the original log function [14] while preserving the convexity of the measure. The DCT space dimension is increased one step further if its TFDC satisfies the inequality

$$C_E(m, Q_m + 1) \geq C_E(m, Q_m) + C_E(m + 1, 1) - \lambda \tag{7}$$

where $C_E(m, Q_m)$ is the TFDC for the current modulation segment and $C_E(m+1, 1)$ is the TFDC for the next single-cycle segment while λ is the difference TFDC threshold, which should be unity for wideband speech with unit-variance residual signal. By increasing the modulation segment length in unit steps, the actual length Q_m is reached when inequality (7) fails to hold.

The baseline for voiced-unvoiced separation is a quite periodic residual whose energy lies completely in the DC modulation coefficient for all harmonic tracks. For a sequence of nonidentical pitch cycles, it is postulated that the lower

$$q_{vm} = \max\{3, \lfloor 0.2Q_m \rfloor\} \tag{8}$$

MT coefficients represent the voiced part while the rest describe the unvoiced part. Therefore, this voiced-unvoiced separation criterion effectively sets $q_{vm} - 1$ as the voicing cutoff frequency for modulation segment m.

Finally, transients are envelope-detected and modeled like the voiced part. As a result, the average modulation segment length is around twelve pitch cycles and may be controlled by the TFDC threshold. Further, the number of MFs in the unvoiced region is noted to be rather larger than that in the voiced region due to the voiced-unvoiced separation criterion.

III. The time-frequency voiced excitation model

The TFVEX model is based on the concept that a voiced model can have lower time resolution [1] and must have greater spectral amplitude resolution. In addition, a high-fidelity representation for the low-frequency PST tracks emphasizes the harmonic structure.

The time resolution for the TFVEX model is the varying modulation segment length since it is a good estimate of the range of local stationarity. An algebraic sign applied to each local standard deviation estimates provides adequate spectral amplitude resolution.

For wideband speech, the $L_h = P_0 = 256$ modulation tracks of same MF are clustered into 9 octave bands that hold $1, 1, 2, 4, \ldots, 128$ tracks, respectively. Low-frequency fidelity is achieved by including the voiced MT coefficients for the lower compound band $B_d = \{l\}_{l=0}^{3}$.

For the set of upper five PST bands $B_u = \bigcup_{b=4}^{8} B_b$, the estimated MT voiced coefficient variance is evaluated as the mean square modulation intensity over its octave band by

$$\left(\hat{\sigma}_{qb}^{(v)}(m)\right)^2 = \frac{1}{2^{b-1}} \sum_{l=2^{b-1}}^{2^b-1} \left(e_{ql}^{(v)}(m)\right)^2 \tag{9}$$

for each upper band $B_b = \left\{2^{b-1}, 2^{b-1}+1, \ldots, 2^b - 1\right\}$ for $b = 4, 5, \ldots, 8$ and MFs in the voiced range $0 \leq q \leq q_{vm} - 1$, where $q_{vm} - 1$ is the highest MF in the voiced region for modulation segment m.

Given the structure above, MT coefficients for the voiced model are retained for the lower band and made equal to the estimated variances affected by the original signs for the upper bands, that is,

$$\hat{e}_{ql}^{(v)}(m) = \begin{cases} e_{ql}^{(v)}(m) & l \in B_d \\ \text{sign}\left(e_{ql}^{(v)}(m)\right)\hat{\sigma}_{qb}^{(v)}(m) & l \in B_b \subset B_u \end{cases} \tag{10}$$

for $l = 0, 1, \ldots, P_0 - 1$ and $0 \leq q \leq q_{vm} - 1$. This modeled modulation segment is built immune to spillover across the voiced-unvoiced boundary by additionally defining

$$\hat{e}_{ql}^{(v)}(m) = 0 \tag{11}$$

for MFs in the unvoiced region $q_{vm} \leq q \leq Q_m - 1$ and harmonic tracks $l = 0, 1, \ldots, P_0 - 1$.

Therefore, the TFVEX model consists of 8 MT coefficients, 5 variances and 248 signs. Packing signs into sets of sixteen 16-bit coded values, it amounts to a total of 29 parameters per voiced MF.

The framework for the voiced model is consistent with the unvoiced model, proposed in [9] and outlined in Section IV. However, TFVEX and TFUNEX differ in their time-frequency footprints and their amplitude resolution within.

IV. The time-frequency unvoiced excitation model

The TFUNEX model is based on the facts that an unvoiced model can have lower spectral resolution [1] and must have greater time resolution. The latter is fulfilled by allowing TFUNEX to have pitch cycle resolution. This is consistent with the observation that the noise in the voiced regions of speech is coherent with the voiced part [8], [5].

The estimated variances for the upper bands are

$$\left(\hat{\sigma}_b^{(u)}(k)\right)^2 = \frac{1}{2^{b-1}} \sum_{l=2^{b-1}}^{2^b-1} \left(c_l^{(u)}(k)\right)^2 \tag{12}$$

for $b = 4, 5, \ldots, 8$. They shape the pitch-synchronous model tracks

$$c_l^{(0)}(k) = \begin{cases} c_l^{(u)}(k) & \text{for } l \in B_d \\ \hat{\sigma}_b^{(u)}(k) x_l(k) & \text{for } l \in B_b \subset B_u \end{cases} \tag{13}$$

for $l = 0, 1, \ldots, P_0 - 1$ and k in modulation segment m, where $x_l(k)$ are generated by independent and identically distributed zero-mean, unit-variance Gaussian processes.

These tracks are modulation-transformed to $e_{ql}^{(0)}(m)$, that lead to the final model, which is obtained as

$$\hat{e}_{ql}^{(u)}(m) = \begin{cases} 0 & \text{for } 0 \leq q \leq q_{vm} - 1 \\ e_{ql}^{(0)}(m) & \text{for } q_{vm} \leq q \leq Q_m - 1 \end{cases} \tag{14}$$

for $l = 0, 1, \ldots, P_0 - 1$, where $q_{vm} - 1$ is the voicing cutoff frequency defined in Section II. In the PST domain, the unvoiced excitation model for modulation segment m is $\hat{c}_l^{(u)}(k)$ for k in modulation segment m and harmonic tracks $l = 0, 1, \ldots, P_0 - 1$.

In total, the TFUNEX model consists of 13 parameters per pitch cycle, which are 8 PST coefficients and 5 variances. Since the number of unvoiced MFs is usually much greater than the number of voiced MFs as noted in closing Section II, the compression effected by TFUNEX is greater than that achieved by TFVEX and the finer time resolution necessary for the unvoiced model is not a heavy penalty. Besides, this explains in part why the unvoiced model has been proposed first [9].

V. The compound time-frequency unvoiced-voiced excitation model

The TFVEX model proposed in Section III and the TFUNEX model described in Section IV can be linearly combined in the modulation domain, giving rise to TFUVEX, the compound unvoiced-voiced model

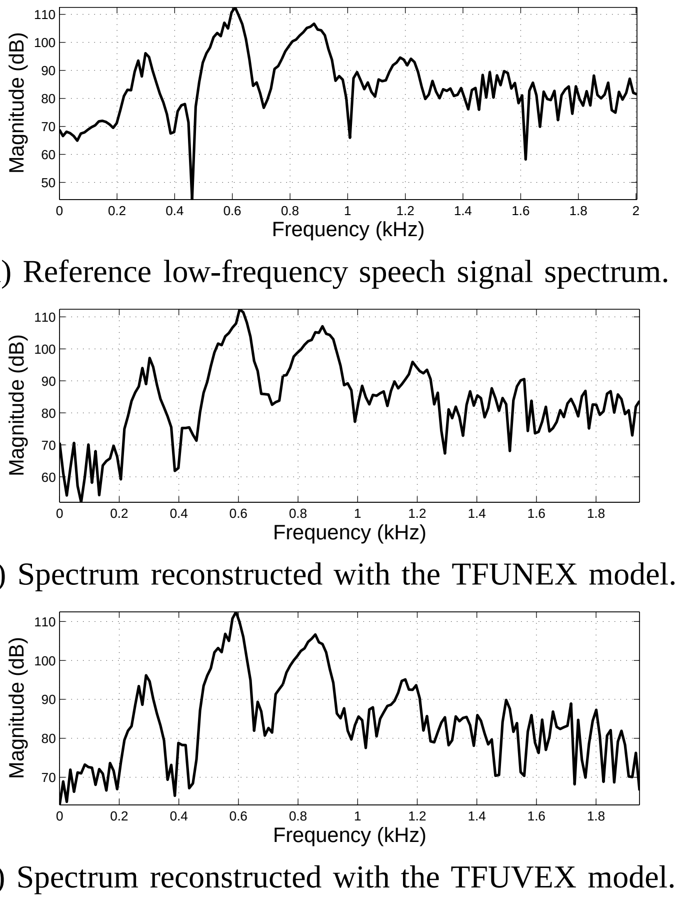
$$\hat{e}_{ql}^{(uv)}(m) = \hat{e}_{ql}^{(v)}(m) + \hat{e}_{ql}^{(u)}(m), \tag{15}$$

for modulation frequencies $0 \leq q \leq Q_m - 1$ in the mth modulation segment and harmonic tracks $0 \leq l \leq P_0 - 1$, where $\hat{e}_{ql}^{(v)}(m)$ is the voiced model modulation coefficient for MF q of the lth harmonic track given by Eq. (10) or (11) and $\hat{e}_{ql}^{(u)}(m)$ is the unvoiced model modulation coefficient for MF q of the lth harmonic track given by Eq. (14). This combination is actually a time-frequency juxtaposition due to the antialiasing operations carried out in the modulation domain in the determination of both models.

It is interesting to note that while the TFUVEX model juxtaposes the models for the voiced and for the unvoiced parts in the modulation domain for flexible processing, the unvoiced model in [8] consists of wavelet transforms around each harmonic peak, which provides for a better recombination.

The spectral fitting performance for TFUVEX is illustrated in Fig. 3(c), where it can be seen to follow rather closely the upper magnitude envelope.

Further, a spectral enhancement has been found beneficial for better low frequency representation and smooth voiced-unvoiced recombination. It is applied to the unvoiced model in the PST domain, where it does not interfere with the periodicity in the signal.



(a) Reference low-frequency speech signal spectrum.

(b) Spectrum reconstructed with the TFUNEX model.

(c) Spectrum reconstructed with the TFUVEX model.

Fig. 3. Low-frequency speech signal spectrum (a) and its reconstruction with ASyS using the time-frequency unvoiced model and original voiced part (b) and using the compound time-frequency unvoiced-voiced model (c).

At first, a highpass filter with cyclic cutoff frequency $f_c = 11/P_0$ is designed by means of the sinc function

$$\mathrm{sinc}(n) = \frac{\sin(\pi n)}{\pi n} \tag{16}$$

to have the impulse response

$$w_h(n) = (1 - 2f_c)\,(-1)^n\,\mathrm{sinc}\left[(1 - 2f_c)\,(n - 32)\right] \tag{17}$$

for $n = 0, 1, \ldots, 64$.

The pitch-synchronous spectral weighting (PSSW) vector plotted in Fig. 4 is defined by sampling the frequency response of the highpass filter as

$$w(l) = W_h\left(e^{j2\pi f_l}\right) \tag{18}$$

at the cyclic frequencies $f_l = l/(2P_0)$, for $l = 0, 1, \ldots, P_0 - 1$.

Finally, the enhanced model coefficients for unvoiced excitation are obtained by applying PSSW as

$$\hat{c}_l^{(u)\prime}(k) = w(l)\hat{c}_l^{(u)}(k) \tag{19}$$

for $l = 0, 1, \ldots, P_0 - 1$.

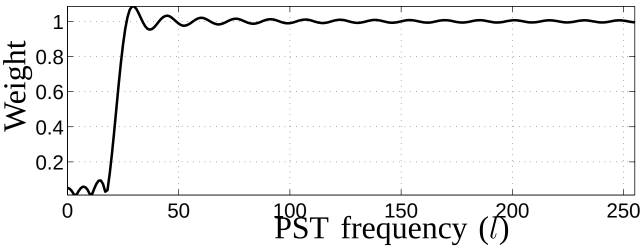


Fig. 4. Pitch-synchronous spectral weights.

VI. Experimental results

For the listening tests, four male and four female utterances were chosen so that each speaker couple belongs to a different dialect region in the TIMIT database[15] for a total length of 448 thousand samples at a sampling rate of 16 kHz. Each signal was synthesized by ASyS in seven test conditions: the analyzed Voiced Part; the modeled Voiced Part plus the analyzed Unvoiced Part – TFVEX; the modeled Voiced Part singled out – TFVEX Alone; the analyzed Voiced Part plus the modeled Unvoiced Part – TFUNEX; the analyzed Voiced Part plus the modeled Unvoiced Part with PSSW emphasis – TFUNEX-PSSW; the modeled Voiced Part combined with the modeled Unvoiced Part – TFUVEX; and the modeled Voiced Part plus the modeled Unvoiced Part with PSSW – TFUVEX-PSSW.

Additionally, five control conditions were prepared for each signal, including the hidden reference and the following four anchors: modulated noise reference unit (MNRU) processed versions at 20 dB, 25 dB, and 30 dB SNR levels; and a 3.5 kHz lowpass-filtered version.

The twelve conditions above were presented to six female and six male listeners as a multi-stimulus with hidden reference and anchors (MUSHRA) test [16] whose results are displayed in Fig. 5.

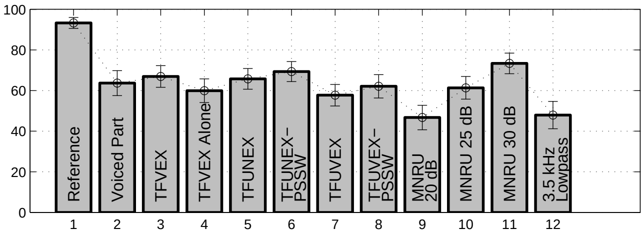


Fig. 5. Mean scores with 95% confidence intervals superimposed in subjective assessment of ASyS provided with TFVEX, TFUNEX and TFUVEX models.

It is striking that the TFVEX Model is rated somewhat above the analyzed Voiced Part and just a bit below when singled out, holding evidence to the high quality of the voiced model. Moreover, all single-model conditions are scored above 25 dB MNRU and the only two-model condition which rises above this level is TFUNEX-PSSW, underlining the distinctive contribution of the unvoiced model to higher fidelity.

Furthermore, the analyzed unvoiced part matters because “TFVEX” is better than “TFVEX Alone”. But it is paradoxical that “TFUVEX” is rated slightly below “TFVEX Alone” as if the addition of the modeled unvoiced part would not improve the modeled voiced part. In fact, the modeled unvoiced part leaks into the voiced region while the analyzed unvoiced part does not. As shown in Section V, this interference is attenuated by means of the PSSW emphasis, which can be seen in Fig. 5 to be an effective measure since “TFUVEX-PSSW” is rated considerably above “TFVEX Alone”.

Besides, all test conditions provide wideband enhancement since their Q values stand significantly above the 20 dB equivalent narrowband version.

The equivalent Q-value is the SNR in dB of the MNRU-impaired signal with the same score as a given condition and provides an inequivocal scale, with a universal meaning [17]. Scores are known to

TABLE I

COMPRESSION RATIOS FOR ANALYSIS BASED UPON TFVEX VOICED MODEL, TFUNEX UNVOICED MODEL AND TFUVEX COMPOUND MODEL ALONG WITH COEFFICIENT AND PARAMETER COUNT BREAKDOWN.

Model	Voiced	Unvoiced	Total	Specific ratio	Overall ratio
Signal	421 k	749 k	1170 k	1.0	1.0
TFVEX	48 k	749 k	797 k	8.8	1.5
TFUNEX	421 k	59 k	480 k	12.7	2.4
TFUVEX	48 k	59 k	107 k	10.9	10.9

be cultural and they may acquire a universal meaning by their equivalent Q-values. Therefore, by using MNRUs at different levels, it is possible to compare the results with any other representation method tested anywhere. For instance, the adaptive multirate wideband (AMR-WB) coder has been rated at an equivalent Q-value of 24 dB when operating at 12.65 kbit/s [18].

The test database has 4569 pitch cycles with 1644 voiced MFs for 1170 thousand raw coefficients. Since TFUNEX represents the unvoiced part of each cycle with 13 parameters, it uses 59 thousand parameters altogether. As for TFVEX, since it requires 29 parameters to represent each voiced MF across all harmonic tracks, it uses 48 thousand parameters altogether. Therefore, as detailed in Table I, the parametric compression ratios are about 9 for TFVEX over the voiced coefficient set and 1.5 overall, 13 for TFUNEX over the unvoiced coefficient set and 2.4 overall and 11 for TFUVEX over all coefficients.

VII. CONCLUSION

Voiced and unvoiced excitation models are proposed for a harmonic time-frequency autoregressive speech analysis-synthesis system. The system features a time-frequency NPR front-end for representation of the prediction residual that is transferred to a modulation domain within modulation segments of varying length based on a time-frequency distribution concentration measure. These modulation segments are the seat for voiced-unvoiced separation so that the voiced part and the unvoiced are appropriately recombined all across the pitch-synchronous time-frequency transform extent. The modulation segment length also defines the time resolution for the TFVEX model while the TFUNEX model has pitch-cycle time resolution and lower amplitude resolution than TFVEX. Both models have octave-band frequency resolution so that they can be conveniently combined into a compound TFUVEX unvoiced-voiced model with spectral weighting. The TFUVEX model generates speech of quality rated above that of speech degraded by 25 dB MNRU while providing a compression of over ten times relative to the NPR time-frequency representation that provides for lower rate parameter coding.

References

[1] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River: Prentice-Hall PTR, 2002, ch. 9.

[2] T. Arai, “Digital Pattern Playback for education in digital signal processing and speech science,” in *Proc. of IEEE Int. Conf. Acoust., Speech, Signal Processing*, Kyoto, 2012, pp. 2769–2772. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6288491

[3] D. Aylln, R. Gil-Pita, P. Jarabo-Amores, M. Rosa-Zurera, and C. Llerena-Aguilar, “Energy-weighted Mean Shift algorithm for speech source separation,” in *Proc. of IEEE Statistical Signal Processing Workshop*, Nice, 2011, pp. 785–788. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5967822

[4] M. Arjona Ramírez and M. Minami, “Split-order linear prediction for segmentation and harmonic spectral modeling,” *IEEE Signal Processing Lett.*, vol. 13, no. 4, pp. 244–247, Apr. 2006. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1605249

[5] Y. Pantazis and Y. Stylianou, “Improving the modeling of the noise part in the harmonic plus noise model of speech,” in *Proc. of IEEE Int. Conf. Acoust., Speech, Signal Processing*, Las Vegas, 2008, pp. 4609–4612. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4518683

[6] M. K. I. Molla, M. A. M. Shikh, and K. Hirose, “Time-frequency representation of audio signals using Hilbert spectrum with effective frequency scaling,” in *Proc. of IEEE Int. Conf. Computer and Information Technology*, vol. 1, Khulna, Bangladesh, 2008, pp. 335–340. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4803077

[7] X. Serra and J. Smith III, “Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition,” *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, Winter 1990. [Online]. Available: http://www.jstor.org/stable/pdfplus/3680788.pdf

[8] P. Polotti and G. Evangelista, “Fractal additive synthesis,” *IEEE Signal Processing Mag.*, vol. 24, no. 2, pp. 105–115, Mar. 2007. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4117933

[9] M. Arjona Ramírez, “Modeling the unvoiced component in the canonical representation of speech,” in *Proc. of DSP 2009 16th International Conference on Digital Signal Processing*, vol. 1, Santorini, Greece, 2009, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5201232

[10] J. Schnitzler, J. Eggers, C. Erdmann, and P. Vary, “Wideband speech coding using forward/backward adaptive prediction with mixed time/frequency domain excitation,” in *Proc. of IEEE Workshop on Speech Coding*, Porvoo, 1999, pp. 4–6. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=781465

[11] H. S. Malvar, “Lapped transforms for efficient transform/subband coding,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 6, pp. 969–978, June 1990. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=56057

[12] N. S. Jayant and P. Noll, *Digital coding of waveforms*. Englewood Cliffs: Prentice-Hall, 1984.

[13] N. Ahmed, T. Natarajan, and K. R. Rao, “Discrete cosine transform,” *IEEE Trans. Comput.*, vol. C-23, no. 1, pp. 90–93, Jan. 1974. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1672377

[14] B. Boashash, Ed., *Time Frequency Signal Analysis and Processing – A comprehensive reference*. Elsevier, 2003, ch. 7.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, “TIMIT acoustic-phonetic continuous speech corpus,” Linguistic Data Consortium, 1993. [Online]. Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

[16] ITU-R, “*Method for the subjective assessment of intermediate quality level of coding systems*,” Recommendation BS.1534-1, Geneva, Jan. 2003. [Online]. Available: http://www.itu.int/rec/R-REC-BS.1534-1-200301-I/e

[17] ITU-T, “*ITU-T Software Tool Library: Software tools for speech and audio coding standardization*,” Recommendation G.191, Geneva, Nov. 2009. [Online]. Available: http://www.itu.int/rec/T-REC-G.191/en

[18] B. Bessette, R. Salami, R. Lefebvre, M. Jelínek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Järvinen, “The adaptive multirate wideband speech codec (AMR-WB),” *IEEE Trans. Speech Audio Processing*, vol. 10, no. 8, pp. 620–636, Nov. 2002, http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1175533.