ELectronics
EXpress

FDMSM robust signal representation for speech mixtures and noise corrupted audio signals

Pejman Mowlaee[a], **Abolghasem Sayadiuan**[b], **and Hamid Sheikhzadeh**[c]

Electrical Engineering Department, Amirkabir University of Technology,
Hafez Avenue, Tehran, 15914, Iran

a) *pmowlaee@ieee.org*
b) *eeas335@cic.aut.ac.ir*
c) *hsheikh@aut.ac.ir*

Abstract: The fixed dimension modified sinusoidal model (FDMSM) was recently proposed as an attractive candidate for compact representation of audio signals in adverse conditions. This paper aims to study the capability of the FDMSM signal representation for analysis and synthesis of speech mixtures as well as noisy audio signals corrupted by highly colored noise of babble and harmonic. Extensive simulation results verified that the FDMSM provides high perceptual quality of the synthesized output signal compared with the conventional harmonic plus noise model (HNM) for both speech mixtures as well as audio signals corrupted by various types of noise.

Keywords: synthesis, HNM, FDMSM, colored noise, speech mixture

Classification: Science and engineering for electronics

References

[1] Y. Stylianou, J. Laroche, and E. Moulines, "High-Quality Speech Modification based on a Harmonic + Noise Model," *Proc. EUROSPEECH*, 1995.

[2] D. Talkin, "Robust pitch tracking," *in Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. Elsevier, 1995.

[3] K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," *ICASSP*, vol. 1, pp. 361–364, 2002.

[4] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, 2004.

[5] M. H. Radfar, A. Sayadiyan, and R. M. Dansereau, "new algorithm for two-speaker pitch tracking in single channel paradigm," *Int. Conf. Signal Processing*, vol. 1, pp. 16–20, Nov. 2006.

[6] P. Mowlaee and A. Sayadiyan, "Sparse Sinusoidal Signal Representation for Speech and Music signals," *CCIS 6*, Springer-Verlag Berlin Heidelberg, pp. 469–476, 2008.

DOI: 10.1587/elex.6.1077
Received April 20, 2009
Accepted July 08, 2009
Published August 10, 2009

[7] M. H. Radfar, A. H. Banihashemi, R. M. Dansereau, and A. Sayadiyan, "Nonlinear minimum square error estimator for mixture-maximization approximation," *Electron. Lett.*, vol. 42, no. 12, pp. 75–76, June 2006.

[8] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, *An audiovisual corpus for speech perception and automatic speech recognition*, JASA 2006. [online] www.dcs.shef.ac.uk/martin/SpeechSeparationChallenge.html

[9] A. Varga and H. Steenneken, "Assessment for automatic speech recognition: Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech commun.*, vol. 12, no. 3, pp. 247–251, 1993.

1 Introduction

In many audio and speech applications it is highly important to deal with speech mixtures and audio signals corrupted with highly correlated noise including babble and harmonic noise. In this respect the lack of an effective signal representation capable for analysis mixed signals is often introduced as a challenging problem in many applications. This difficulty is mainly due to that many pitch-dependent analysis methods including harmonic plus noise model (HNM) [1], are not capable of analyzing mixed signals since the performance of the state-of-the-art pitch estimation algorithms [2, 3, 4] may severely degrade by introducing large errors for mixtures [5]. The purpose of this study is to investigate the capability of the FDMSM signal representation for a wide range of audible signals including speech mixtures, songs, and noise corrupted audio signals. The results are compared with those obtained by HNM as a pitch-synchronous benchmark algorithm.

2 FDMSM signal model

Recently in [6], it was shown that a compact representation is possible by a novel subband-based sinusoidal model called FDMSM (see Fig. 1). The FDMSM as a whole consists of two parts, analysis and synthesis stage as depicted in Fig. 1. The analysis stage is composed of two parts namely, (1) signal segmentation followed by taking the Discrete-time Fourier Transform (DFT), and (2) filtering followed by peak picking. The aim behind spectral smoothing is to remove the frequency components within $f < 62.5\,\text{Hz}$ and $f > 3840\,\text{Hz}$ from the magnitude spectrum prior to peak picking due to low (50 or 60 Hz) and high frequency harmful effects (assuming sampling frequency $f_s = 8\,\text{kHz}$). Finally, peak picking is performed.

As the number of peaks obtained at the end of the first stage differs from one segment to another due to pitch dependence of speech signals at various frames, a feature selection is employed. The aim of feature selection is to apply an appropriate transformation by which the variable feature dimension is made fixed at frames. However, the required transformation should preserve the synthesized quality approximately the same as the original signal in terms of perception.

DOI: 10.1587/elex.6.1077
Received April 20, 2009
Accepted July 08, 2009
Published August 10, 2009

The subband transformation is accomplished by designing several subbands. Each subband is characterized with two parameters centre frequency in addition to the related bandwidth. The number of subbands (FDMSM model order) is,

$$\widehat{\mathrm{M}} = \left\lceil \frac{\Omega(\omega_{\widehat{\mathrm{M}}}) - \Omega(\omega_1)}{\Delta\Omega} \right\rceil \tag{1}$$

where $\widehat{\mathrm{M}}$ is the number of mel-scaled subbands (FDMSM model order), $\Omega(\omega_1)$, $\Omega(\omega_{\widehat{\mathrm{M}}})$ are mel-scale start and end frequencies, respectively while $\Delta\Omega$ is the frequency interval. Next, a search is made within each band to determine the peak candidates. Finally, the decision is made on these peak candidates whether to pick or discard. The output of feature selection is a set of sinusoids called the FDMSM feature parameters. Next, the FDMSM parameters obtained in analysis stage are converted in time-domain shown as $v_i^k(\mathrm{n}), \mathrm{i} \in [1, \widehat{\mathrm{M}}]$ where i denotes the ith subband and k is the frame index. Sinusoidals are put together by a weighted overlap-add (WOLA) to reconstruct the synthesized signal, $\hat{\mathrm{s}}^{\mathrm{k}}(\mathrm{n})$.

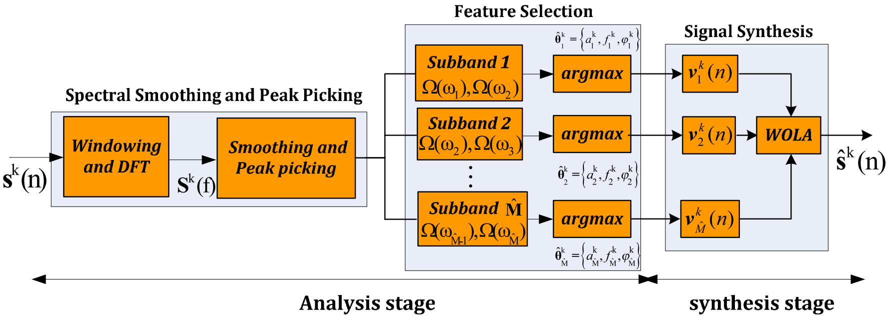

Feature Selection
Signal Synthesis
Spectral Smoothing and Peak Picking
$\mathbf{s}^k(n)$
Windowing and DFT
$S^k(f)$
Smoothing and Peak picking
Subband 1 $\Omega(\omega_1),\Omega(\omega_2)$
Subband 2 $\Omega(\omega_2),\Omega(\omega_3)$
Subband $\hat{M}$ $\Omega(\omega_{\hat{M}-1}),\Omega(\omega_{\hat{M}})$
argmax
argmax
argmax
$\hat{\theta}_1^k=\{a_1^k,f_1^k,\varphi_1^k\}$
$\hat{\theta}_2^k=\{a_2^k,f_2^k,\varphi_2^k\}$
$\hat{\theta}_{\hat{M}}^k=\{a_{\hat{M}}^k,f_{\hat{M}}^k,\varphi_{\hat{M}}^k\}$
$v_1^k(n)$
$v_2^k(n)$
$v_{\hat{M}}^k(n)$
WOLA
$\hat{\mathbf{s}}^k(n)$
Analysis stage
synthesis stage

Fig. 1. FDMSM: consists of four parts; (a) analysis stage, 2) smoothing plus peak picking, 3) FDMSM feature selection, and taking maximum peak at each mel-scaled subband, and (b) synthesis: WOLA for signal reconstruction.

2.1 Problem formulation and signal model

Consider a frame of a speech signal composed of a set of impulses corrupted by white noise of constant power as,

$$\boldsymbol{x}(n) = \underbrace{\mathrm{Re}\left\{\sum_{\mathrm{i}=1}^{\mathrm{M}} \alpha_i \mathrm{e}^{\mathrm{j}2\pi \mathrm{n} f_i}\right\}}_{\boldsymbol{s}(n)} + \mathbf{w}(n), \qquad 0 \le n \le \mathrm{N}-1 \tag{2}$$

where N is the analysis time-window length in samples, n is the time-sample index and $\mathbf{w} = [\mathrm{w}(n)\ \mathrm{w}(n+1)\ \ldots\ \mathrm{w}(n+\mathrm{N}-1)]^{\mathrm{T}}$ is the white noise, α_i and f_i are the sinusoidal amplitude and frequency, respectively. Note we assume that $\boldsymbol{s}(n)$ is composed of M sinusoidals. According to (2), an observed signal $\boldsymbol{x}(n)$ can be decomposed into two signals, namely, a speech signal denoted by $\boldsymbol{s}(n)$ plus noisy components shown by $\mathbf{w}(n)$. Each sinusoidal frequency f_i is then defined as a sinusoidal frequency vector as,

© IEICE 2009
DOI: 10.1587/elex.6.1077
Received April 20, 2009
Accepted July 08, 2009
Published August 10, 2009

$\boldsymbol{v}_i^k = [1 \quad e^{j2\pi f_i} \quad \dots \quad e^{j2\pi f_i(N-1)}]^{\mathrm{H}}, i \in [1, \mathrm{M}]$ where $\boldsymbol{v}_i^k$ is the *ith* frequency vector of dimension $\mathrm{N} \times 1$ composed of length-N DFT vector, index i denotes the ith subband with frequency f_i (as depicted in the synthesis stage in Fig. 1) The FDMSM synthesized signal is shown as,

$$\hat{s}^k(n) = \sum_{i=1}^{\widehat{\mathrm{M}}} \hat{a}_l^k \cos(2\pi \hat{f}_i^k n + \hat{\varphi}_i^k) \tag{3}$$

where $\hat{a}_i^k$, $\hat{f}_i^k$, $\hat{\varphi}_i^k$ are the amplitude, frequency and phase parameters for the *kth* frame estimated by the FDMSM, $\hat{\mathrm{s}}^{\mathrm{k}}(\mathrm{n})$ is the synthesized speech of the kth frame, $\widehat{\mathrm{M}}$ the number of sinusoidals in the FDMSM. In a matrix form all the sinusoidal frequencies obtained by the FDMSM shown by $\boldsymbol{v}_i^k, 1 \leq i \leq \widehat{\mathrm{M}}$ are represented by, $\widehat{\mathbf{V}}_{\boldsymbol{s}}^{\mathrm{k}} = [\hat{\boldsymbol{v}}_1^k \; \hat{\boldsymbol{v}}_2^k \; \dots \; \hat{\boldsymbol{v}}_{\widehat{\mathrm{M}}}^k]^H, 1 \leq i \leq \widehat{\mathrm{M}}$ where $\widehat{\mathbf{V}}_{\boldsymbol{s}}^{\mathrm{k}}$ is an $\widehat{\mathrm{M}} \times \mathrm{N}$ matrix, the columns of $\widehat{\mathbf{V}}_{\boldsymbol{s}}^{\mathrm{k}}$ are the FDMSM vectors composed of frequencies $\hat{f}_i$ with $i = 0, \dots, \widehat{\mathrm{M}}$. The FDMSM signal representation is,

$$\hat{s}^{\mathrm{k}} = [\widehat{\mathrm{A}}_1^{\mathrm{k}} \dots \widehat{\mathrm{A}}_{\widehat{\mathrm{M}}}^{\mathrm{k}}][\hat{\boldsymbol{v}}_1^k \dots \hat{\boldsymbol{v}}_{\widehat{\mathrm{M}}}^k]^H = \widehat{\mathbf{A}}_1 \widehat{\mathbf{V}}_1^{\mathrm{k,H}} \tag{4}$$

where $\widehat{\mathbf{A}}_1 = [\widehat{\mathrm{A}}_{1,1}^{\mathrm{k}} \dots \widehat{\mathrm{A}}_{1,\widehat{\mathrm{M}}}^{\mathrm{k}}]$ is a $1 \times \widehat{\mathrm{M}}$ vector composed of the amplitude parameters obtained from applying FDMSM to the kth frame of the observed signal, and $\boldsymbol{s}^k$ is $1 \times \mathrm{N}$ waveform and denotes complex sinusoidal amplitude vector at the kth frame.

2.2 FDMSM max approximation within subbands

In this section we provide a mathematical analysis for the FDMSM signal representation. Consider two speech frames of different speakers, $\mathrm{s}_1^{\mathrm{k}}(\mathrm{n})$ and $\mathrm{s}_2^{\mathrm{k}}(\mathrm{n})$ at the *kth* frame index. Generally, any observed speech signal is composed of two different parts, harmonic and noise components. Each speaker speech segment can then be expressed as a product sum of FDMSM plus a model error term as,

$$\mathrm{s}_{\mathrm{j}}^{\mathrm{k}}(\mathrm{n}) = \widehat{\mathbf{A}}_j \mathbf{V}_{\mathrm{j,s}}^{\mathrm{H}} = [\widehat{\mathrm{A}}_{\mathrm{j},1}^{\mathrm{k}} \dots \widehat{\mathrm{A}}_{\mathrm{j},\widehat{\mathrm{M}}}^{\mathrm{k}}] \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ e^{j2\pi(N-1)\hat{f}_{j,1}} & \cdots & e^{j2\pi(N-1)\hat{f}_{j,\widehat{\mathrm{M}}}} \end{bmatrix}^H + \varepsilon_j^k(n) \tag{5}$$

$$\mathrm{s}_{\mathrm{j}}^{\mathrm{k}}(n) = [\widehat{\mathrm{A}}_{\mathrm{j},1}^{\mathrm{k}} \dots \widehat{\mathrm{A}}_{\mathrm{j},\widehat{\mathrm{M}}}^{\mathrm{k}}][\hat{\boldsymbol{v}}_{j,1}^k \dots \hat{\boldsymbol{v}}_{j,\widehat{\mathrm{M}}}^k]^H + \varepsilon_j^k(n) \tag{6}$$

where $\widehat{\mathbf{V}}_{\mathrm{j,s}}^{\mathrm{k}} = [\hat{\boldsymbol{v}}_{j,1}^k \dots \hat{\boldsymbol{v}}_{j,\widehat{\mathrm{M}}}^k]^H, j \in [1, 2]$ denotes the FDMSM frequency vectors, $\varepsilon_j(n)$ is the model error of the FDMSM signal representation for the jth speaker in speech mixture of $\mathrm{z}(n) = \mathrm{s}_1^{\mathrm{k}}(n) + \mathrm{s}_2^{\mathrm{k}}(n)$ and $\widehat{\mathrm{M}}$ is the model order of FDMSM. $f_{j,k}$, and $\widehat{\mathrm{A}}_{\mathrm{j,i}}^{\mathrm{k}}$ denote the FDMSM amplitude for *ith* subband for the *jth* speaker in the mixture. Note the components for each speaker is in fact the maximum peak found in the DFT spectrum at each frame subband. We define subband spectrum of the *jth* speaker signal at the *kth* frame as $\mathrm{s}_{\mathrm{j,i}}^{\mathrm{k}}(\Omega)$ with $\mathrm{j} \in [1, 2]$ and $\Omega \in [\Omega_{\mathrm{i}-1}, \Omega_{\mathrm{i}}]$ where Ω indicates the mel-scale frequency. For each FDMSM subband we have: $\Omega(\omega_{\mathrm{i}}) < \Omega_{\mathrm{i}}(\omega) < \Omega(\omega_{\mathrm{i}+1}), i = 1, \dots, \widehat{\mathrm{M}} - 1$.

© IEICE 2009
DOI: 10.1587/elex.6.1077
Received April 20, 2009
Accepted July 08, 2009
Published August 10, 2009

As shown in Fig. 1. The FDMSM signal representation could also be reformulated as,

$$\begin{aligned}\mathbf{s}_1^{\mathrm{k}}(\mathrm{n}) &= \sum_{\mathrm{i}=1}^{\widehat{\mathrm{M}}} A_{1,i}^{k} \mathrm{e}^{\mathrm{j}2\pi \mathrm{n}\Omega_{\mathrm{i}}} + \varepsilon_1^{k}(n) \\ &= [\widehat{\mathrm{A}}_{1,1}^{\mathrm{k}} \ldots \widehat{\mathrm{A}}_{1,\widehat{\mathrm{M}}}^{\mathrm{k}}][\hat{\boldsymbol{v}}_{1,1}^{k}(\Omega_{\mathrm{i}}) \ldots \hat{\boldsymbol{v}}_{1,\widehat{\mathrm{M}}}^{k}(\Omega_{\mathrm{i}})] + \varepsilon_1^{k}(n) \\ &= \widehat{\mathbf{A}}\widehat{\mathbf{V}}_{\mathbf{s}}^{\mathbf{H}}(\Omega) + \varepsilon_1^{k}(n) \end{aligned} \tag{7}$$

where $\boldsymbol{v}^k(\Omega_{\mathrm{i}}) = [1 \quad e^{j2\pi\Omega_i} \quad \ldots \quad e^{j2\pi\Omega_i(N-1)}]^{\mathrm{H}}$, $1 \leq i \leq \widehat{\mathrm{M}}$ denotes the FDMSM frequency vectors obtained in mel-scaled subbands (warped frequency domain), $\widehat{\mathbf{V}}_{\mathrm{s}}(\Omega)$ denotes the matrix composed of all sinusoids obtained from the FDMSM in the mel-scaled frequencies $\hat{\boldsymbol{v}}(\Omega_{\mathrm{i}}), 1 \leq i \leq \widehat{\mathrm{M}}$. $\mathrm{A}_{1,i}$ $\mathrm{i} \in [1, \widehat{\mathrm{M}}]$ indicates the magnitude spectrum of the FDMSM peaks at each mel-scaled subband. These peaks are chosen by selecting the peak with the highest amplitude in the logarithmic spectrum at each subband. Hence by replacing $A_{j,i} = \mathrm{Max}\{\mathrm{s}_{\mathrm{j,i}}^{\mathrm{k}}(\Omega)\}$ with $j = 1, 2$ the speaker index and $1 \leq i \leq \widehat{\mathrm{M}}$ as subband index we have,

$$\begin{aligned}\hat{\mathbf{s}}_{\mathrm{j}}^{\mathrm{k}}(\mathrm{n}) &= \sum_{\mathrm{i}=1}^{\widehat{\mathrm{M}}} \mathrm{Max}\{\mathrm{s}_{\mathrm{j,i}}^{\mathrm{k}}(\Omega_{\mathrm{j}})\}\mathrm{e}^{\mathrm{j}2\pi\mathrm{n}\Omega_{\mathrm{j,i}}} \\ &= [\widehat{\mathrm{A}}_{\mathrm{j},1}^{\mathrm{k}} \ldots \widehat{\mathrm{A}}_{\mathrm{j},\widehat{\mathrm{M}}}^{\mathrm{k}}][\hat{\boldsymbol{v}}_{j,1}^{k}(\Omega_{\mathrm{j,i}}) \ldots \hat{\boldsymbol{v}}_{j,\widehat{\mathrm{M}}}^{k}(\Omega_{\mathrm{j,i}})]^{H}, \ \mathrm{j} = 1, 2 \end{aligned} \tag{8}$$

where $\Omega_{\mathrm{j,i}}$ denotes the mel-scaled frequency of the *ith* subband spectrum for the *jth* speaker obtained by taking maximum from $\mathrm{s}_{\mathrm{j,i}}^{\mathrm{k}}(\Omega)$ of the jth speaker signal. Eq. (8) can also be interpreted as a subband maximum approximation. As the FDMSM seeks for the maximum element of the logarithmic spectrum per subbands, the FDMSM peaks can be expressed as

$$\{\widehat{\mathrm{A}}_{\mathrm{j},1}^{\mathrm{k}} \ldots \widehat{\mathrm{A}}_{\mathrm{j},\widehat{\mathrm{M}}}^{\mathrm{k}}\} = \mathrm{Max}\{\log|\mathrm{s}_{\mathrm{j},1}^{\mathrm{k}}(\Omega_{\mathrm{j}})|, \log|\mathrm{s}_{\mathrm{j},2}^{\mathrm{k}}(\Omega_{\mathrm{j}})|, \ldots, \log|\mathrm{s}_{\mathrm{j},\widehat{\mathrm{M}}}^{\mathrm{k}}(\Omega_{\mathrm{j}})|\}, \ \mathrm{j} \in [1, 2] \tag{9}$$

assuming that phase values are modeled as a uniform distribution, the maximum approximation is recently proven as a nonlinear MMSE estimator of the log spectra of the underlying speech signals as shown in [7]. Hence, Eq. (9) indicates that the FDMSM parameter estimation is a MMSE estimator at each subband and FDMSM representation solves the estimation problem per subbands as,

$$\hat{\mathrm{z}}^{\mathrm{k}}(\mathrm{n}) = \sum_{\mathrm{i}=1}^{\widehat{\mathrm{M}}} \mathrm{Max}\{\mathrm{s}_{1,\mathrm{i}}^{\mathrm{k}}(\Omega_{\mathrm{i}}) + \mathrm{s}_{2,\mathrm{i}}^{\mathrm{k}}(\Omega_{\mathrm{i}})\}\hat{\boldsymbol{v}}_{\mathrm{z,i}}^{\mathrm{k}}(\mathrm{argmax}\{\mathrm{s}_{\mathrm{z,i}}^{\mathrm{k}}(\Omega_{\mathrm{i}})\}) = \sum_{\mathrm{i}=1}^{\widehat{\mathrm{M}}} \widehat{\mathrm{A}}_{\mathrm{z,i}}^{\mathrm{k}}\hat{\boldsymbol{v}}_{\mathrm{z,i}}^{\mathrm{k}}(\Omega_{\mathrm{i}}) \tag{10}$$

3 Simulation results

The evaluation results for the FDMSM are compared with those obtained by HNM as benchmark. The results are averaged over ten mixtures each composed of a clean speech signal and a colored noise of a particular type. The speech utterances are chosen from [8] the noise signals are taken from the Noisex [9].

DOI: 10.1587/elex.6.1077
Received April 20, 2009
Accepted July 08, 2009
Published August 10, 2009

3.1 Audio signals corrupted by babble noise

The Perceptual Evaluation of Speech Quality (PESQ) is used to evaluate the performance for speech and music signals corrupted with babble noise at SNR=10 dB. The results, summarized in Table I test 1, show a reasonable PESQ value of more than 3.5 for the noisy speech. It is also observed that in average the FDMSM outperforms HNM approximately with 1.5 to 2 advantage of PESQ score. From Table I it is observed that FDMSM reaches at a fixed value of PESQ as $\widehat{\mathrm{M}}$ grows.

3.2 Signals corrupted by harmonic noise

Experiment is conducted for various input signals including single speaker speech signal, speech mixture composed of two speakers and music signals contaminated with a harmonic noise. The harmonic noise has a fundamental frequency of 300 Hz and there are 10 harmonics in its spectrum. The objective evaluation results are summarized in Table I test 2. The PESQ results in Table I test 2 indicate that the FDMSM achieves reasonable performances of 1.2 advantage over HNM. The most notable result is related to the category of the male speaker signal where the FDMSM provides an improvement of 1.6 in PESQ compared with HNM.

3.3 Speech mixture

Objective assessment for the speakers in terms of SSNR (Segmental Signal-to-Noise Ratio), PESQ, Log-Likelihood ratio (LLR), and Weighted Spectral Slope (WSS) versus the FDMSM model order, $\widehat{\mathrm{M}}$ and HNM are presented in Table II. Test 1 is composed of a speech mixture formed by employing speakers from [8] mixed at SSR=0 dB. From the table, it is observed that the FDMSM requires $33 < \widehat{\mathrm{M}} < 40$ to reconstruct the speech mixture. Also note any change in $\widehat{\mathrm{M}}$ in this range has little effect on the objective evaluations. In comparison, HNM results in poor performance for speech mixture as shown in Table II test 1. In this case, FDMSM results in a higher perceptual quality compared with HNM.

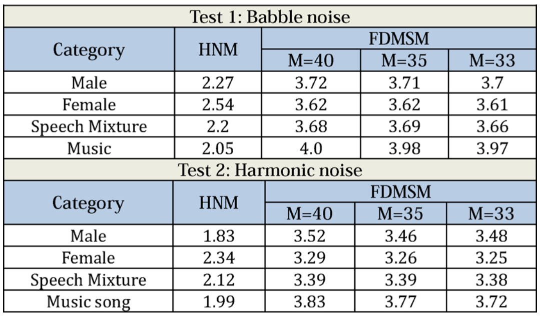
For a song music signal in Table I test 2, it was observed that the FDMSM could achieve a PESQ higher than 4 leading to a synthesized signal indistinguishable from the original. The poor performance of HNM for mixtures is due to erroneous pitch estimation since polyphonic signal may easily lead to errors in HNM. This is also true for speech mixtures where two pitch frequencies may easily overlap for different speakers. Another explanation for HNM performance degradation may be that too many pitch candidates are found and as a result the parameter estimation of HNM is erroneous. Subjective evaluations showed that the difference between the synthesized signal by the FDMSM and the original signal is mostly negligible while HNM synthesized signals result in poor performance. The MOS results averages to 3.6 for the FDMSM compared with 2 obtained by HNM. The wave files of the original and FDMSM reconstructed signals are downloadable from: http://kom.aau.dk/˜pmb/IEICE.

DOI: 10.1587/elex.6.1077
Received April 20, 2009
Accepted July 08, 2009
Published August 10, 2009

4 Conclusion

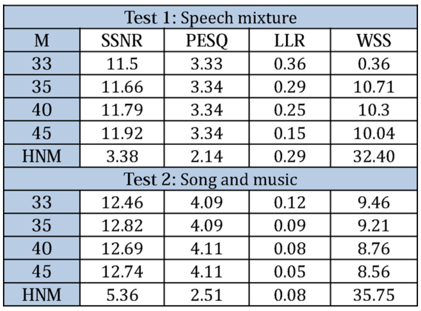
In this paper, the capability of recently proposed FDMSM signal representation was investigated for speech mixtures of speakers and audio signals corrupted by noise. The performance of the FDMSM reconstructed signal was compared with the conventional HNM as a pitch-dependent benchmark model. Simulation results show that FDMSM provides an attractive candidate with high perceptual quality for compact representation of speech mixtures and noise corrupted audio signals.

Table I. Quality assessment in PESQ for babble and harmonic noise at SNR=10 dB.



Test 1: Babble noise				
Category	HNM	FDMSM		
		M=40	M=35	M=33
Male	2.27	3.72	3.71	3.7
Female	2.54	3.62	3.62	3.61
Speech Mixture	2.2	3.68	3.69	3.66
Music	2.05	4.0	3.98	3.97
Test 2: Harmonic noise				
Category	HNM	FDMSM		
		M=40	M=35	M=33
Male	1.83	3.52	3.46	3.48
Female	2.34	3.29	3.26	3.25
Speech Mixture	2.12	3.39	3.39	3.38
Music song	1.99	3.83	3.77	3.72

Table II. Objective assessment for speech mixture and music song versus model order.



Test 1: Speech mixture				
M	SSNR	PESQ	LLR	WSS
33	11.5	3.33	0.36	0.36
35	11.66	3.34	0.29	10.71
40	11.79	3.34	0.25	10.3
45	11.92	3.34	0.15	10.04
HNM	3.38	2.14	0.29	32.40
Test 2: Song and music				
33	12.46	4.09	0.12	9.46
35	12.82	4.09	0.09	9.21
40	12.69	4.11	0.08	8.76
45	12.74	4.11	0.05	8.56
HNM	5.36	2.51	0.08	35.75

© IEICE 2009
DOI: 10.1587/elex.6.1077
Received April 20, 2009
Accepted July 08, 2009
Published August 10, 2009