TIME-SCALING OF SPEECH AND MUSIC USING INDEPENDENT SUBSPACE ANALYSIS

R. Muralishankar[1], *Lakshmish N. Kaushik*[2] *and A. G. Ramakrishnan*[2]

[1]INRS-EMT (Telecommunications), University of Quebec, Montreal, Canada.
[2]Department of Electrical Engineering, Indian Institute of Science, Bangalore-560012, INDIA.
murali@inrs-emt.uquebec.ca, ramkiag@ee.iisc.ernet.in, lakshmish@ragashri.ee.iisc.ernet.in

ABSTRACT

We propose a new technique for modifying the time-scale of speech and music using Independent Subspace Analysis (ISA). To carry out ISA, the single channel mixture signal is converted to a time-frequency representation such as spectrogram. The spectrogram is generated by taking Hartley or Wavelet transform on overlapped frames of speech or music. We do dimensionality reduction of the autocorrelated original spectrogram using singular value decomposition. Then, we use Independent component analysis to get unmixing matrix using JadeICA algorithm [1]. It is then assumed that the overall spectrogram results from the superposition of a number of unknown statistically independent spectrograms. By using unmixing matrix, independent sources such as temporal amplitude envelopes and frequency weights can be extracted from the spectrogram. Time-scaling of speech and music is carried out by resampling the independent temporal amplitude envelopes. We then multiply the independent frequency weights with time-scaled temporal amplitude envelopes. We Sum these independent spectrograms and take inverse Hartely or wavelet transform of the sum spectrogram. The reconstructed time-domain signal is overlap-added to get the time-scaled signal. The quality of the time-scaled speech and music has been analyzed using Modified Bark Spectral Distortion(MBSD) [2]. From the MBSD score, one can infer that the time-scaled signal is less distorted.

1. INTRODUCTION

Time-scale modification of speech or music refers to processing performed on signals that changes the perceived rate of articulation without affecting the pitch or intelligibility of the signals. Such modification can be categorized into two classes: time-scale compression (or speed-up) that increases the rate of articulation; and time-scale expansion (or slow-down) that decreases the rate of articulation. Time and frequency are important cues for the auditory system. For example, they relate to the loudness and intensity of a signal. Psychoacoustics also relies on time and frequency information. For instance, increasing the time base of consonants during speech increases intelligibility and comprehension. Additionally, the temporal structure of speech itself is largely determined by the periodic clousure of the glottis [3]. The relation between consonant time base and comprehension leads to a variety of applications for time-scaling. Time-scale compression is used in faster listening of messages recorded on answering machines, voice mail systems, and other information services. On the otherhand, the goal of slow-down (time-scale expansion) is to aid in comprehension or dictation of rapidly spoken speech segments with important information, such as an address or phone number. In the cellular phone industry, prolonging relevant segments of speech in real-time might lead to clearer conversations. In the case of music, time-scaling is also used as a tool for composing music.

Several algorithms have been developed to achieve time-scale modification based on the inherent structure of the audio signal. Time-domain techniques rely on the periodic nature of speech, while analysis/synthesis techniques exploit redundancies in the signal to reduce the speech waveform to a limited set of time varying parameters. Time-domain techniques operate by inserting or deleting segments of audio signal, which can result in discontinuities in the transition between inserted or deleted segments. The Time-domain harmonic scaling (TDHS) algorithm [4] determines the local pitch by employing multiple correlations of signal segments. A triangular windowing function is aligned with the pitch periods and the resulting segments are added such that pitch periods are inserted or deleted to create a time-scale modified signal. The algorithm requires exact pitch determination to operate successfully. It provides good quality in the class of low complexity time-domain algorithms. There are a few alternatives to this method, such as Synchronized Overlap-Add (SOLA), which was originally proposed by Roucos and Wilgus [5], and Waveform Similarity Overlap-Add (WSOLA), proposed by Verhelst and Roelands [6]. These techniques have low complexity and operate in the time-domain, but do not rely on pitch tracking. As these methods use fixed window lengths and fixed windowing intervals, they have advantages for real-time implementation and are being used in both speech and music time-scaling.

In the fequency-domain techinques, Phase vocoder (voice coder) [7] is the most popular method for time and frequency scaling. This is mainly because of its efficient implementation using FFT [8]. It separates a signal into its instantaneous phase and amplitude components, which could then be used to modify time-scale or pitch-scale of a signal. Portnoff [9] proposed a vocoder based time-scale modification scheme using short-time Fourier transform. Present day time-scaling techniques use wavelets instead of constant bandwidth spectral analysis of FFT, to best match the non-uniform characteristics of human auditory systems.

Our method uses fixed frame length to generate spectrograms of speech and music signals. Similarly, frame lengths were also chosen for music signals. However, from our observation, for getting a good time-scaled speech, one needs to choose frame length depending on approximate pitch period of the signal under consideration. Real transform has been used to generate the spectrogram to avoid handling of phase at the reconstruction stage. We reduce the dimension of the spectrogram followed by Independent component analysis (ICA). To achieve the required time-scaling, we resample the independent temporal envelopes. Finally, we add all the time-scaled independent spectrograms and resynthesise to get the time-scaled signal.

2. INDEPENDENT SUBSPACE ANALYSIS (ISA)

Casey's innovation in ISA [10] was to take a mono signal (that cannot ordinarily be unmixed directly using ICA) and perform a change of basis operation before employing canonical ICA techniques. Based on redundancy reduction techniques, it represents sound sources as low dimensional independent subspaces in the time-frequency plane. ISA makes a number of assumptions about the nature of the signal and the sound sources present in the signal. The single channel speech mixture is assumed to be a sum of 'p' unknown independent sources,

$$s(n) = \sum_{q=1}^{p} s_q(n) \tag{1}$$

Taking Hartley transform of the signal and using the 'k' coefficients for 'm' frames yields a spectrogram of the signal, S of dimension $k \times m$, where k is the number of frequency channels, and m is the number of time slices. From this, it can be seen that each column of S contains a vector which represents the frequency spectrum at time l, with $1 \leq l \leq m$. Similarly each row can be seen as the evolution of frequency channel over time, with $1 \leq i \leq k$. It is assumed that the overall spectrogram S results from the superposition of 'l' unknown independent spectrograms S_j. As the superposition of spectrograms is a linear operation in the time-frequency plane this yields:

$$S = \sum_{j=1}^{l} S_j \tag{2}$$

It is then assumed that each of the S_j can be uniquely represented by the outer product of an invariant frequency basis function f_j, and a corresponding invariant amplitude envelope or weight function 't_j' which describes the variations in amplitude of the frequency basis function over time. This yields

$$S_j = f_j t_j^T \tag{3}$$

Summing S_j yields

$$S = \sum_{j=1}^{l} f_j t_j^T \tag{4}$$

In practice, the assumption that the frequency basis functions are stationary means that no change in pitch can occur within the spectrogram. Casey and Westner [10] overcame this assumption by breaking the signal into smaller blocks within which the pitch can be considered stationary.

The independent basis functions correspond to features of the independent sources, and each source is composed of a number of these independent basis functions. The basis functions that compose a sound source form a low-dimensional subspace that represents the source. The basis functions are selected based upon capturing maximum variance present in the spectrogram (optimal information for source separation). Once the low-dimensional subspaces have been identified, the independent sources can be resynthesized. In our approach, we do resampling of the amplitude envelope or weighing function t_j before resynthesizing to achieve the required time-scaling.

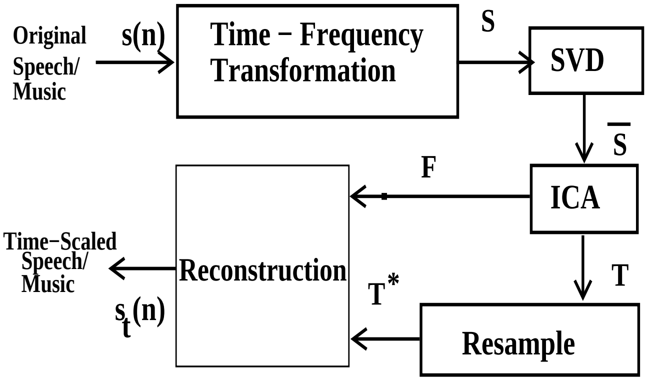

Fig. 1. Block diagram of Time-scaling using ISA.

3. TIME-SCALING USING ISA

Figure 1 shows the block diagram of time-scale modification using ISA. A description of the preprocessing and the calculation of independent frequency basis function and amplitude envelope is presented in detail in the following subsections.

3.1. Preprocessing

The speech data is divided into a number of frames with an overlap equal to half the frame length. Here, the frame-length has been chosen based on twice the average pitch period of the speech signal. It is windowed using a hamming window and mapped to the spectral domain using real transforms such as Discrete cosine transform (DCT), Discrete sine transform (DST) or Hartley transform. We have also used sub-band based approach to map the speech data into the spectral domain. We get the spectrogram after the mapping, where it has 'k' frequency bins and 'm' frames (time slices).

3.2. Singular value decomposition

Consider a transposed spectrogram as the matrix S^T. Its singular value decomposition (SVD) is given by

$$S^T = UDV^T \tag{5}$$

The application of SVD is equivalent to the eigenvalue decomposition of the covariance matrix S^T. Standard SVD algorithms return a diagonal matrix D of singular values in decreasing order and two orthogonal matrices U& V^T. Matrix $U = (u_1, ...u_m)$, also referred to as the row basis, holds the left singular vectors, which is equal to the eigenvectors of SS^T. Matrix $V = (v_1, ...v_n)$ also referred to as the column basis, holds the right singular vectors equal to the eigenvectors of $S^T S$. The singular vectors are linearly independent and therefore provide the orthonormal basis for a rational transform in the directions of the principal components.

3.3. Reduction of dimensionality

The SVD orders the basis vectors according to the size of their singular values. The singular values represent the standard deviations of the principal components of S. These standard deviations are proportional to the amount of information contained in the corresponding principal components. A maximally informative subspace of the input data S is obtained by applying the following procedure.

A linear transformation G is calculated as given below, where $\overline{D}$ is a submatrix consisting of the upper 'd' rows of D.

$$G = \overline{D} V^T \tag{6}$$

The transformation matrix G is multiplied with the spectrogram S, yielding a representation $\overline{S}$ of reduced rank and maximally informative orientation:

$$\overline{S} = GS \tag{7}$$

The number 'd' of retained dimensions is a meaningful parameter of the spectrogram. However, from our observations, a limited number (30 to 70) of dimensions is sufficient for getting good resynthesized speech. Dimensions fewer than this lead to an incomplete decomposition and hence poor resynthesized speech, while more dimensions give no reasonable improvement in the perceived resynthesized speech. Higher dimensions also increase the computational load.

3.4. Independent component analysis (ICA)

Source separation model is a transformation, where the observations x are obtained by a multiplication of the source signals s by an unknown mixing matrix A. The reduced rank spectrum $\overline{S}$ can be interpreted as an observation matrix, where each column is regarded as realizations of a single observation. In this work, the Jade-ICA algorithm [1] is applied for the estimation of A. It minimizes higher order correlations by joint approximate diagonalization of eigen matrices of cross cumulant tensors. The estimated matrix A is used to calculate the independent components. Its pseudo-inverse A^{-1} represents the unmixing matrix, using which the independent sources can be extracted. The reduced rank spectrogram $\overline{S}$ is modified to obtain the independent temporal amplitude envelopes T.

$$T = A^{-1}\overline{S} \tag{8}$$

The independent frequency weights F are estimated by the folowing expression and a subsequent pseudo-inversion.

$$F^{-1} = A^{-1}G \tag{9}$$

The independent spectrograms are computed by multiplying one column of F with the corresponding row of T,

$$S_c = F_{u,c}T_{c,v} \tag{10}$$

where $u = 1, ..., k$, $v = 1, ..., m$ and $c = 1, ..., d$.

3.5. Time-scaling

We now resample T depending on the time-scale factor, i.e., for factors > 1, we time-stretch the input signal and for factors < 1, we time-compress it. We denote the resampled temporal amplitude envelopes as $T^\star$. Finally, independent spectrograms are computed (after resampling) by multiplying one column of F with the corresponding row of $T^\star$, as shown below.

$$S_c^\star = F_{u,c}T_{c,v^\star}^\star \tag{11}$$

where $u = 1, ..., k$, $v^\star = 1, ..., m^\star$ and $c = 1, ..., d$. For $m^\star > m$, reconstructed speech is expanded in time and for $m^\star < m$, reconstructed speech is compressed.

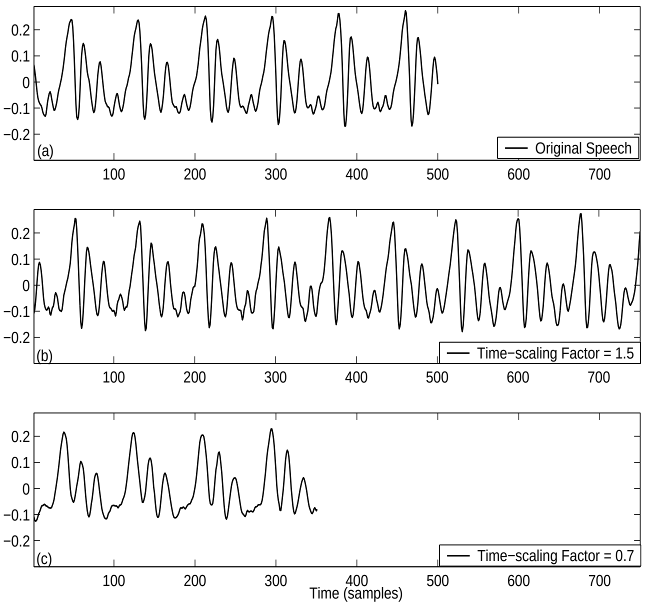

Fig. 2. Time-scaling using ISA. (a) Few frames of the original signal. (b) Few frames of the signal time-scaled by a factor of 1.5. (c) Few frames of the signal time-scaled by a factor of 0.7.

3.5.1. Sub-band ISA based Time-scaling

Sub-band based approach removes the restriction of fixed resolution and introduces multi-resolution in mapping from time-domain to time-frequency domain. We call this as sub-band spectrogram. To generate sub-band spectrogram, we use Biorthogonal wavelet, because it exhibits the property of linear phase, which is needed for signal and image reconstruction. Once we get the sub-band spectrogram, we follow the same steps as explained in previous subsections, to achieve time-scaling.

3.6. Reconstruction

After resampling of independent temporal amplitude envelopes, we sum all the independent spectrograms and then inverse transform the sum-spectrogram, to get time-domain signal, which is the resultant of overlapped and time-scaled version of the input signal. The time-domain signal is overlapped and added with the same frame-length and shift. This removes the windowing effect in the time-scaled signal.

4. RESULTS AND DISCUSSION

To evaluate the abilities of the present approach, we tested on spoken sentences from different speakers. These sentences were recorded using SM-58 microphone under less noisy conditions. As discussed previously, we chose the frame-length approximately equal to twice the average pitch period of the signal under consideration. Figure 2 shows few frames of time-expanded and compressed signals along with few frames of the original signal (Fig. 2(a)). In Fig. 2(b) and 2(c), we have shown few frames of ISA based time-scaled signals for the factors 1.5 and 0.7, respectively. We can see small temporal deviation of the time-scaled speech compared to original speech and with the pitch being intact, as shown in Fig. 2. Figure

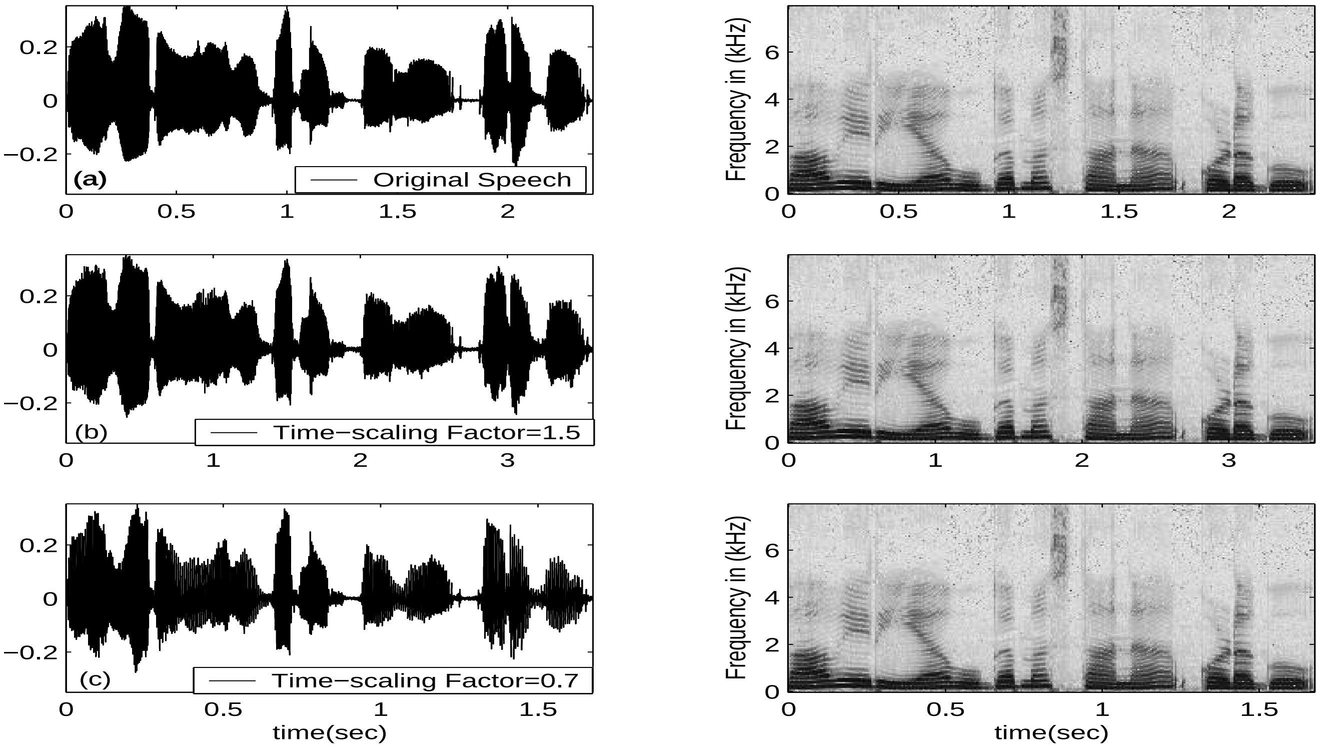

Fig. 3. Time-scaled speech signals (left panel) and corresponding spectrograms (right panel). (a) original speech signal /kaveriya ugamasthana kodagu/. (b) Time-scaled speech signal (scaling factor = 1.5). (c) Time-scaled speech signal (scaling factor = 0.7)

3 shows the time-scaled speech signals and corresponding spectrograms, respectively. One can see the close matching of the spectrogram between original and time-scaled speech signals. We tested our approach to time-scale music signals too. Music signal samples were taken from the recorded *Carnatic classical music* performances from different artists. Figure 4 shows the time-scaled music signals and corresponding spectrograms, respectively. We can notice the close matching of the spectrograms of original and time-scaled music signals.

To measure the quality of time-scaled speech, we used an objective measure that correlates well with the subjective quality measure. Among various objective measures, we use Modified Bark Spectral Distortion (MBSD) [2]. This estimates speech distortion in the loudness domain, taking into the account the noise masking threshold in order to include only audible distortions in the calculation of the distortion measure. Its performance improvement over Bark Spectral Distortion (BSD) has been presented in [2]. BSD measure is the average squared Euclidean distance of estimated loudness of the original and the coded utterances. Even though the conventional BSD measure showed a relatively high correlation with mean opinion score (MOS), there are areas of possible improvement. Motivated by the transform coding of audio signals, which uses the noise masking threshold, the MBSD measure has incorporated this concept of noise masking threshold into the conventional BSD measure, where any distortion below the threshold is not included for the calculation of distortion. This new addition of the noise masking threshold replaces the empirically derived distortion threshold value used in the conventional BSD [2]. Since MBSD compares the distorted speech to the original speech, its performance would be sensitive to the temporal misalignment. So a synchronization algorithm based on loudness domain is applied prior to performing the MBSD [11]. Upon applying MBSD on our time-scaled speech and music, the results were encouraging. The distortion values are close to zero, indicating good quality and less distortion in the time-scaled signals. Tables 1 and 2 show the MBSD scores for speech and music, respectively. We can see little increase in the distortion when we use fixed transform (Hartley) to generate spectrogram compared to sub-band approach using wavelets. Little increase in the distortion for the time-scaled speech and music can be seen, as the factors move away from 1 in both directions.

5. CONCLUSION

We presented a new method for time-scale modification using ISA. In this method, independent temporal amplitude envelopes have been resampled to achieve the required time-scaling. The advantage in our approach lies in the fact that we need to get independent temporal amplitude envelopes and frequency weights only once for a given speech or music signal. The MBSD measure indicates negligible distortion in the speech and music signals, time-scaled using our method.

6. REFERENCES

[1] “Jade algorithm for independent component analysis,” http://www.tsi.enst.fr/icacentral/algos.html.

[2] W Yang, M Benbouchta, and R Yantorno, “Performance of the modified bark spectral distortion as an objective speech quality measure,” *ICASSP-98*, vol. 1, pp. 541–544, 1998.

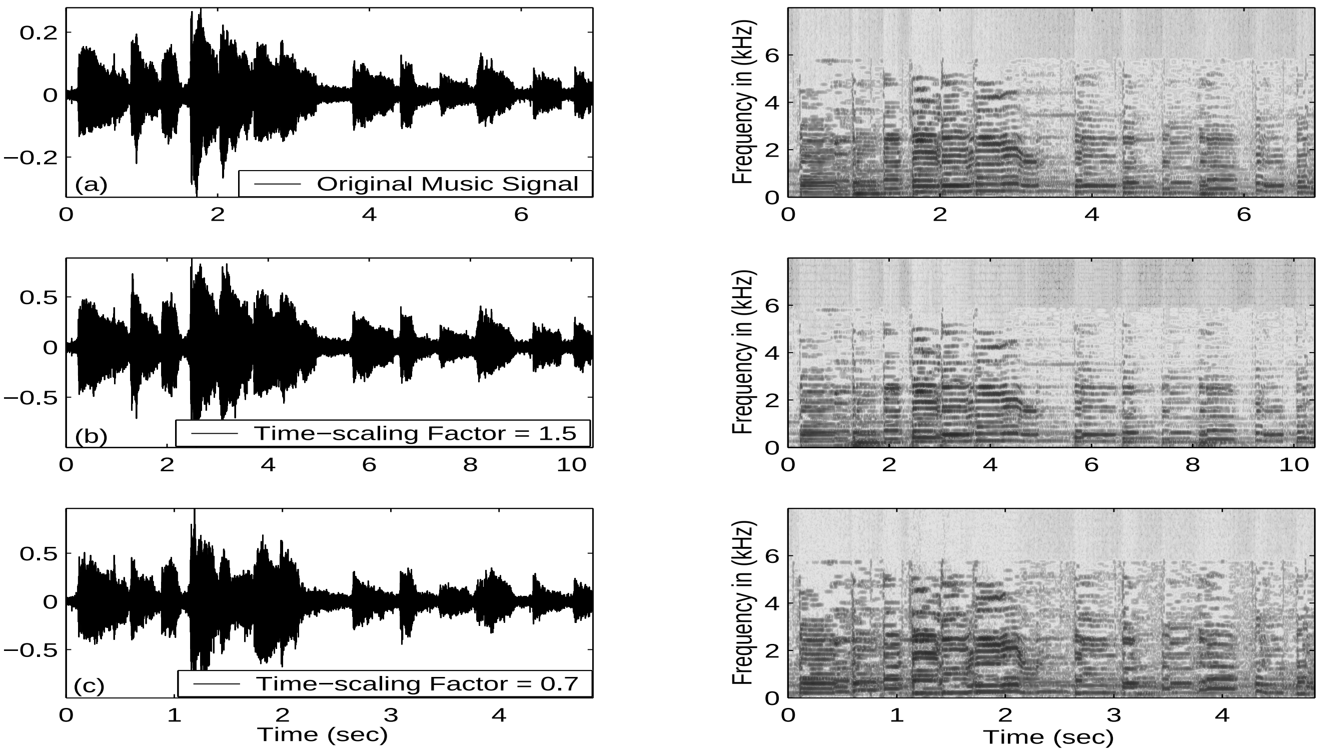

Fig. 4. Time-scaled music signals (left panel) and corresponding spectrograms (right panel). (a) original music signal (Instrument used is Veena). (b) Time-scaled music signal (scaling factor = 1.5). (c) Time-scaled music signal (scaling factor = 0.7).

Table 1. MBSD scores for Time-scaled speech

	MBSD Scores (10^{-4})	
Time-scaling factor	Hartley Transform	Wavelet Transform
0.5	2.82	2.44
0.8	3.94	3.64
1.1	2.07	1.97
1.4	2.76	2.41
1.7	4.00	3.50
2	4.92	4.42

Table 2. MBSD scores for Time-scaled music

	MBSD Scores (10^{-4})	
Time-scaling factor	Hartley Transform	Wavelet Transform
0.5	6.02	5.23
0.8	5.34	4.27
1.1	5.00	4.18
1.4	6.74	6.21
1.7	8.21	7.64
2	9.42	8.68

[3] *Phase Interpolation methods for Pitch and Time-scale modification of Voiced Speech*, vol. 18 of *Proceedings of Institute of Acoustics*, Nov. 1996.

[4] David Malah, "Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals," *IEEE Trans. on ASSP*, vol. 27, pp. 121–133, Apr. 1979.

[5] W Verhelst and M Roelands, "An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," *ICASSP-93*, pp. 554–557, 1993.

[6] *An Overlap-Add Technique based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech*, ICASSP, 1993.

[7] J L Flanagan and R L Golden, "Phase vocoders," The bell system technical journal, Nov. 1966.

[8] M R Portnoff, "Implementation of the digital phase vocoder using fast fourier transform," *IEEE Trans. on ASSP*, vol. 24, no. 3, pp. 243–248, June 1976.

[9] M R Portnoff, "Time-scale modification of speech based on short-time fourier analysis," *IEEE Trans. on ASSP*, vol. 29, no. 3, pp. 374–390, 1981.

[10] *Separation of Mixed Audio Sources by Subspace Analysis*, Proceedings of ICMC, 2000.

[11] M Benbouchta, "A waveform synchronization algorithm in the context of objective measure of speech quality," M.S. thesis, Temple University, Philadelphia, PA, 1998.