TIME-SCALING OF SPEECH USING INDEPENDENT SUBSPACE ANALYSIS

R. Muralishankar, A. G. Ramakrishnan and Lakshmish N. Kaushik

Department of Electrical Engineering, Indian Institute of Science, Bangalore-560012, INDIA.
(murali,lakshmish)@ragashri.ee.iisc.ernet.in, ramkiag@ee.iisc.ernet.in

ABSTRACT

We propose a new technique for modifying the time-scale of speech using Independent Subspace Analysis (ISA). To carry out ISA, the single channel mixture signal is converted to a time-frequency representation such as spectrogram. Here, the spectrogram is generated by taking Hartley or Wavelet transform on overlapped frames of speech. 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 [5]. 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 is carried out by resampling the independent temporal amplitude envelopes. We then obtain time-scaled independent spectrograms after multiplying the independent frequency weights with time-scaled temporal amplitude envelopes. Summing all these independent spectrograms and taking inverse Hartely or wavelet transform of the sum spectrogram to reconstruct and overlap-add the reconstructed time-domain signal to get the time-scaled speech. The quality of the time-scaled speech has been analyzed using Modified Bark Spectral Distortion(MBSD) [6]. From the MBSD score, one can infer that the time-scaled signal is less distorted.

1. INTRODUCTION

Time-scale modification of speech refers to processing performed on speech signals that changes the perceived rate of articulation without affecting the pitch or intelligibility of the speech. Such modification can be categorized into two classes: time-scale compression (or speed-up) which increases the rate of articulation; and time-scale expansion (or slow-down) which decreases the rate of articulation. Traditional uses of time-scale modification allow for 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.

Several algorithms have been developed to achieve time-scale modification based on the inherent structure of the speech 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 speech signal, which can result in discontinuities in the transition between inserted or deleted segments. The (Time-domain harmonic scaling) TDHS algorithm [1] 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 [2], and Waveform Similarity Overlap-Add (WSOLA), proposed by Verhelst and Roelands [3]. 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.

Our method uses fixed frame length to generate spectrogram of the speech signal. 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 and 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 [4] was to take a mono signal (that ordinarily cannot be unmixed directly using ICA) and perform a change of basis operation before employing cannonical 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 on the signal and using the 'k' coefficients for 'm' slices 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 j, with $1 \leq i \leq m$. Similarly each row can be seen as the evolution of frequency channel over time, with $1 \leq j \leq k$. It is

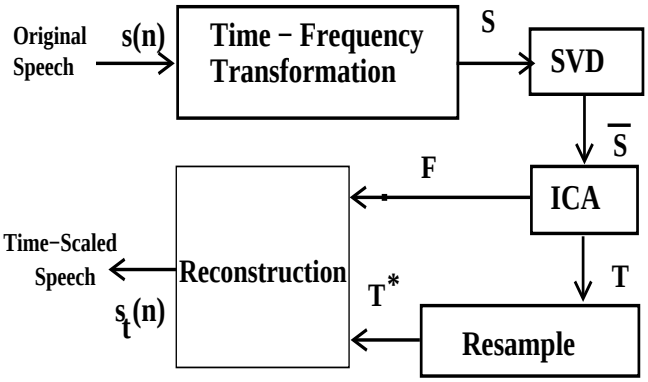

Fig. 1. Block diagram of Time-scaling using ISA.

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 weighting 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 [4] 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 in otherwords 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 weighting function t_j before resynthesizing to achieve the required time-scaling.

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 overlapped frames with an overlapped interval 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), Hartley transform etc., 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^TS. The singular vectors are linearly independent and ther efore provide the orthonormal basis for a rational transform into 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 following procedure.

A linear transformation G is calculated according to the eq. 6. 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 as given in eq. 7.

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

The number 'd' of retained dimensions is a meaningful parameter of the spectrogram. However, from our observations a limited amount of 30 upto 70 dimensions is sufficient for getting good resynthesized speech. Fewer dimensions lead to an incomplete decomposition and hence poor resynthesized speech, while more dimensions give no reasonable improvement in the perceived resynthesized speech. Higher dimensions 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 JadeICA algorithm [5] 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, by which the independent sources can be extracted. Employing eq. 8 modification of the independent temporal amplitude envelopes T are obtained from the reduced rank spectrogram $\overline{S}$.

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

The estimation of the independent frequency weights F is achieved by eq. 9 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

After obtaining independent frequency weights F and independent temporal amplitude envelopes T from the reduced rank spectrogram $\overline{S}$, we then resample T depending on the time-scale factor, i.e., for factors > 1, result in time-stretching of the input signal and for factors < 1, result in time-compression. 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 in eq. 11.

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

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.

3.5.1. Sub-band ISA based Time-scaling

Sub-band based approach removes the restriction of fixed resolution and introduce 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 instead of normally used Daubechies 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 similar steps 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 later inverse transforming of the sum-spectrogram, we get time-domain signal which is resultant of overlapped and time-scaled version of the input signal. The time-domain signal is overlapped and add with the same frame-length and shift. This removes the windowing effect in the time-scaled signal.

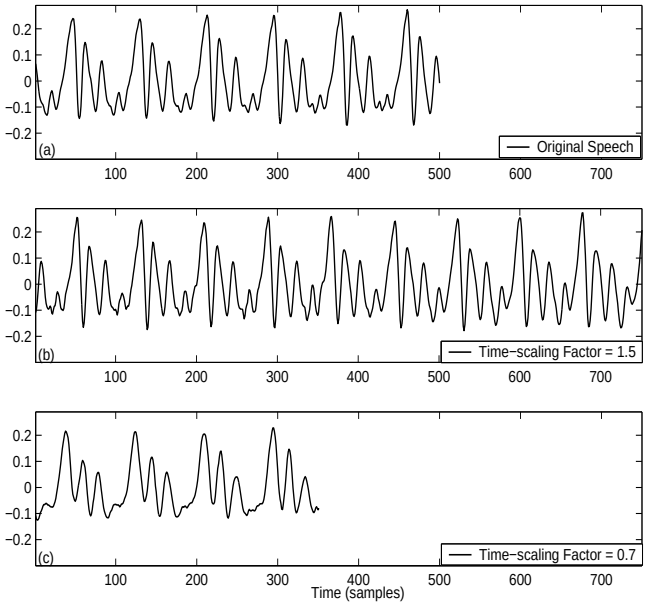

Fig. 2. Time-scaling using ISA(a) Few frames of the original signal. (b) Few frames of the signal time-scaled by a factor of 1.5. (c) Few frames of the signal time-scaled by a factor of 0.7.

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 choose 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 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 compare to original speech and with the pitch being intact, as shown in Fig. 2. Figure 3 shows the time-scaled signals and corresponding spectrogram towards right side, respectively. One can see the close matching of the spectrogram between original and time-scaled signals.

To measure the quality of time-scaled speech, we used objective measure that correlates well with the subjective quality measure. Among various objective measures, we use Modified Bark Spectral Distortion (MBSD) [6]. 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 [6]. 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

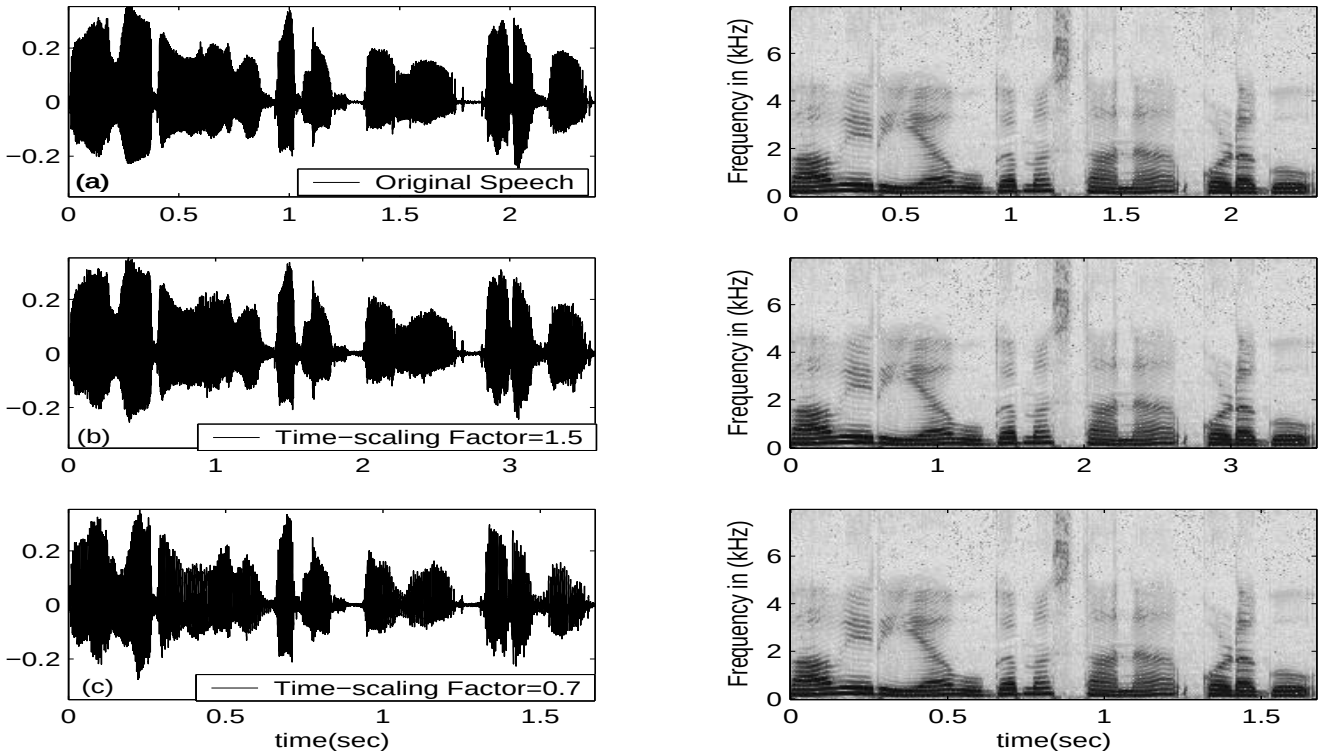

Fig. 3. Time-scaled signals (left panel) and corresponding spectrograms (right panel). (a) original speech signal /kaveriya ugamasthana kodagu/. (b) Time-scaled signal (scaling factor = 1.5) of (a). (c) Time-scaled signal (scaling factor = 0.7) of (a).

MBSD measure has incorporated this concept of noise masking threshold into the conventional BSD measure, where any distortion below the noise masking 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 [6]. Since the 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 [7]. Upon applying MBSD on our time-scaled speech, the results were encouraging interms of the distortion values close to zero, indicating good quality and less distortion in the time-scaled speech (as shown in Table 1).

Table 1. MBSD scores for Time-scaled speech

Time-scaling factor	MBSD score (10^{-5})
0.6	8.6214
0.7	7.8323
0.8	7.3194
1.2	8.9955
1.5	12.2300
1.8	13.9330
2.0	17.2440

5. CONCLUSION

We presented here a new method for time-scale modification using ISA. In this method, resampling of independent temporal amplitude envelope has been done 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 signal; the required time-scaling is obtained than by resampling of independent amplitude envelopes. The MBSD measure indicates negligible distortion in the time-scaled speech using our method.

6. REFERENCES

[1] David Malah, 'Time-Domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Speech Signals', *IEEE Trans. on ASSP*, Vol. ASSP-27, No. 2, pp. 121-133, April 1979.

[2] S. Roucos and A. Wilgus, 'High quality Time-Scale Modification of Speech', in *Proc. ICASSP-85*, pp. 236-239, 1985.

[3] W. Verhelst and M. Roelands, 'An Overlap-Add Technique based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech', in *Proc. ICASSP-93*, pp. 554-557, 1993.

[4] M. A. Casey and A. Westner, 'Separation of Mixed Audio Sources by Subspace Analysis', *Proc. of ICMC-2000*, pp. 154-161, 2000.

[5] Jade algorithm for ICA. http://www.tsi.enst.fr/icacentral/algos.html

[6] 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, Seattle, 1998.

[7] M. Benbouchta, 'A Waveform synchronization algorithm in the context of objective measure of speech quality', Master Thesis, Electrical and Computer Engineering Department, Temple University, Philadelphia, PA, 1998.