Restoring the missing features of the corrupted speech using linear interpolation methods

Taha H. Rassem, Nasrin M. Makbol, Ali Muttaleb Hasan, Siti Syazni Mohd Zaki, and P. N. Girija

Citation: AIP Conference Proceedings **1891**, 020119 (2017); doi: 10.1063/1.5005452
View online: http://dx.doi.org/10.1063/1.5005452
View Table of Contents: http://aip.scitation.org/toc/apc/1891/1
Published by the American Institute of Physics

Restoring the missing features of the corrupted speech using linear interpolation methods

Taha H. Rassem[1, a)], Nasrin M. Makbol[2] , Ali Muttaleb Hasan[1], Siti Syazni Mohd Zaki[1] , P. N. Girija[3]

[1]*Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang, Kuantan, Malaysia.*
[2]*School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Penang, Malaysia.*
[3]*School of Computer and Information Sciences, UoH, Hyderabad, T.S, India.*

[a)]Corresponding author: tahahussein@ump.edu.my,

Abstract. One of the main challenges in the Automatic Speech Recognition (ASR) is the noise. The performance of the ASR system reduces significantly if the speech is corrupted by noise. In spectrogram representation of a speech signal, after deleting low Signal to Noise Ratio (SNR) elements, the incomplete spectrogram is obtained. In this case, the speech recognizer should make modifications to the spectrogram in order to restore the missing elements, which is one direction. In another direction, speech recognizer should be able to restore the missing elements due to deleting low SNR elements before performing the recognition. This is can be done using different spectrogram reconstruction methods. In this paper, the geometrical spectrogram reconstruction methods suggested by some researchers are implemented as a toolbox. In these geometrical reconstruction methods, the linear interpolation along time or frequency methods are used to predict the missing elements between adjacent observed elements in the spectrogram. Moreover, a new linear interpolation method using time and frequency together is presented. The CMU Sphinx III software is used in the experiments to test the performance of the linear interpolation reconstruction method. The experiments are done under different conditions such as different lengths of the window and different lengths of utterances. Speech corpus consists of 20 males and 20 females; each one has two different utterances are used in the experiments. As a result, 80% recognition accuracy is achieved with 25% SNR ratio.

INTRODUCTION

Automatic Speech Recognition (ASR) can be considered as a statistical pattern classification problem [1-4]. The input of the ASR system is a sequence of feature vectors that will be extracted after applying windowing on the input speech. From the training database (corpus of speech), the ASR system able to learn the feature vectors distribution and which sound that belong to each feature vector. From the clean training speech, the distribution of these vectors will be learnt. The Speech recognition accuracy degrades when the speech is corrupted by a noise. This is because of the mismatch between the distribution of feature vectors of the noisy speech and the distribution of the feature vectors of the clean speech (training speech). One solution to this problem is to retrain the system with the same level of noise to avoid the mismatching. This solution does not help either the noise is stationary or non-stationary. Many researchers [1-3] have been suggested some methods for noise reduction to improve the performance of ASR systems. Many of proposed methods depend on reducing and manipulating the effect of the noise in the earlier stages of the ASR systems. Reducing the effect of the noise can be called compensation. The compensation methods are classified into two approaches. In the first one, the test data need some manipulation to compensate the effect of noise to get a similar distribution of the training data, this approach is called as data - compensation. In the latter approach, the training data is modified to be similar to a test data to reduce the mismatching between the distributions of the training and testing data, this approach is called as classifier-compensation. There are many classical data compensation methods like Codeword Dependent Cepstral

The 2nd International Conference on Applied Science and Technology 2017 (ICAST'17)
AIP Conf. Proc. 1891, 020119-1–020119-6; https://doi.org/10.1063/1.5005452
Published by AIP Publishing. 978-0-7354-1573-7/$30.00

Normalization (CDCN), Vector Taylor Series (VTS), Spectral Subtraction (SS) and Wiener filtering (WF). Moreover, there are many of classical classifier compensation methods such as Parallel Model Combination (PMC) and Model Composition (MC)[5]. The classifier-compensation and data-compensation can be performed and improve the recognition accuracy well only with the medium and low level of noise. So, there is a difficultly to handle a low signal to noise ratio.

Speech is transformed into the time-frequency domain. It can be represented as spectrographic images with two axes of the image represent time and frequency, respectively. The pixel value of each element in the spectrographic image represents the energy of the signal in that time-frequency location [2, 6]. When the noise has been added to the speech, the noise will affect each pixel according to the energy in this location because the noise in each pixel differs from other pixels. Missing feature methods are one of the compensation methods. In the missing features methods, the locations in the time-frequency spectrogram are marked by binary mask based on the local SNR. The recognition using the incomplete spectrogram can be done; this method has been called as incomplete spectrogram method. Incomplete spectrogram methods have been shown remarkably robust to high levels of noise corruption [2]. On the other hand, special techniques are used to reconstruct the erased pixels first. Then, the recognition will be on the new reconstructed complete spectrogram. Incomplete spectrogram methods are called the classifier compensation methods while the reconstructed spectrogram methods as data compensation methods. The missing feature methods showed a successful performance for compensation the noise in the corrupted speech either in stationary and non-stationary noise conditions. In this paper, we are targeting the missing feature methods. The linear interpolation methods that are considered as geometrical methods are implemented as a toolbox for reconstruction the missing features of the corrupted speech. Linear interpolation along time and frequency have been tested with different windows sizes and steps for speech corpus consists of 20 males and 20 females; each one has two different utterances. Moreover, a new interpolation method is presented by considering both of the time axis and frequency axes together in the interpolation process to restore the missing features. All the experiments in this paper are evaluated using Sphinx CMU software. This paper is organised as follows. Section 2 explains some of the spectrogram reconstruction methods. The experimental setup and results are presented in Section 3. Finally, the conclusion of the paper is given in Section 4.

SPECTROGRAM RECONSTRUCTION METHODS

In the data compensation methods, the missing data should be estimated before performing the recognition, the methods that will be used for reconstructing the missing features are referred as spectrogram reconstruction methods. This section covers important spectrogram reconstruction methods that can be used to predict and restore the missing data from the available data such as the geometrical structure of speech spectra or some statistical information needed from the clean speech that will be available to achieve good reconstruction. The main aim of applying these methods is not only for good reconstruction but also for increasing the recognition accuracy with the reconstructed spectrograms. In the geometrical method, the reconstruction is based on the data are available in the observed regions within the spectrogram. All geometrical information inside the spectrogram will be used to reconstruct the spectrogram-missing regions [7-10]. The spectrogram features show continuity in all frequency and time axes. Based on that, it can expect that the spectrogram frequency components of the spectral vectors show statistical dependencies with the components of the same vector and with the other vectors components within the same spectrogram. In the statistical reconstruction methods, the statistical dependencies between the various components of within the same spectrogram can be easily learned from the uncorrupted speech. Statistical relations learned are called “vector statistics”, which shows the distribution of the spectral vectors as well as the statistical relationship among the various frequency components within spectral vectors, or “covariance statistics”, which shows the statistical relationship between the components of different vectors in the spectrogram. These statistical relations are used for reconstruction the missing features. In this paper, the geometrical methods will explain and the results of these methods will discuss in different conditions.

Geometrical Reconstruction Methods

They are two types of geometrical reconstruction methods. These two methods are working on the same concept, which is the interpolation. In these methods, the missing element will be reconstructed by doing the interpolation process within the spectrogram between the adjacent observed elements. These adjacent elements could be along either the frequency axis or the time axis of the spectrogram. The interpolation used could be a simple linear

interpolation or non-linear both along time and frequency [1-4]. Generally, the interpolation is used for estimating the value lies between some other values, so this method can be used for estimating any missing element by interpolating between its neighbours[9].

If we have any sequence of numbers such as $s[1], s[2], s[3].....,s[M]$, and we have samples in the interval $[l_1, l_2]$ are unknown or missing. The estimation of the missing values using the linear interpolation can be obtained by drawing a straight line between the nearest known neighbors, $s[l_1 - 1]$ $s[l_2 + 1]$ and reading the estimating values of $s[l_1]$ through $s[l_2]$ off the line. This operation can be expressed mathematically as follows:

$$s\hat{}[l] = s[l_1 - 1] + \frac{(s[l_2 + 1] - s[l_1 - 1])(l - l_2 + 1)}{l_2 - l_1 + 2} \tag{1}$$

Where $l_1 \le l \le l_2$

Linear Interpolation along Frequency

In this method, the interpolation will be done on each spectral vector to estimate the missing elements in this vector based on its observed frequency elements. Mathematically, in each spectral vector l, the estimated value of $s(l,k)$ where the k lies between the frequency components $[k_1, k_2]$ equals

$$s\hat{}(l,k) = s(l,k_1 - 1) + \frac{s(l,k_2 + 1) - s(l,k_1 - 1)}{k_2 - k_1 + 2}(k - k_1 + 1) \tag{2}$$

$s(l,k_2 + 1)$ and $s(l,k_1 - 1)$ elements are the neighbors of $s\hat{}(l,k)$ in the spectral vector l. The interpolation can be done when the missing element k lies between two known elements $[k_1, k_2]$. When the missing elements are at the spectrogram boundaries, it is difficult to estimate these missing elements. For example, if $s(l,k)$, $k_1 \le k \le k_2$ is missing and $k_1 = 1$ or $k_2 = K$, it is difficult to estimate this element by interpolation along frequency. This is because the spectral vector has observed components on one side only of the missing elements. In this case, we can only estimate the missing elements by applying the linear extrapolation of the two closest observed elements. For example, if the closest observed components of the vector are $s(l,k_3)$ and $s(l,k_4)$, the mathematical expression of the missing elements can be described as follows:

$$s\hat{}(l,k) = s(l,k_3) + \frac{s(l,k_4) - s(l,k_3)}{k_4 - k_3}(k - k_3) \tag{3}$$

Linear Interpolation along Time

In interpolation along time method, the estimation process will be performed between the same frequency elements in adjacent spectral vectors. In this case, the sequence of the points that will be used for interpolation process would be a single slice form (time slice) of the spectrogram, i.e., parallel to the time axis. Hence, if the k^{th} frequency component in vector $[l_1, l_2]$, i.e. $s(l,k)$, $l_1 \le l \le l_2$ were missing, the estimation of these missing values could be obtained as follows:

$$s\hat{}(l,k) = s(l_1 - 1,k) + \frac{s(l_2 + 1,k) - s(l_1 - 1,k)}{l_2 - l_1 + 2}(l - l_1 + 1) \quad (4)$$

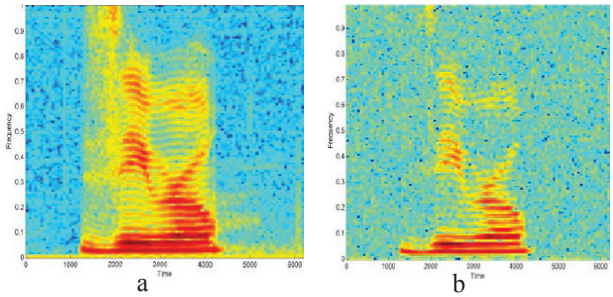
Based on Equation (4), the missing features are estimated by interpolation within the same frequency components and using different time instants. However, if $s(l,k)$, $l_1 \le l \le l_2$ are missing and $l_1 = 1$ or $l_1 = N$, we cannot do the interpolation along time where all the observed values of frequency component k will be in one side of the missing segment. In this case, the missing elements can be estimated by linear extrapolation using the closest observed elements to the missing components in the time slice. The Equation (5) shows the mathematical expression of the linear extrapolation between two closest observed elements in the time slice.

$$s\hat{}(l,k) = s(l_3,k) + \frac{s(l_4,k) - s(l_3,k)}{l_4 - l_3}(l - l_3) \quad (5)$$

Reconstruct the missing elements depend mainly on remaining elements remaining after adding the noise or randomly drop the elements [10]. Therefore, all the methods presented in the paper are considered as local reconstruction methods. This is because, the missing features are obtained from the same data i.e., spectrogram itself without any other information or external information sources.

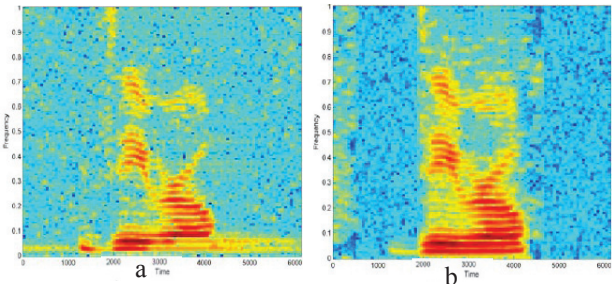
EXPERIMENTAL RESULTS

The goal of reconstruction of the missing features is not only to get the good reconstruction of the spectrogram but also to achieve good recognition accuracy from reconstructed signals. In this section, the results of some reconstructed spectrogram methods are mentioned. The SPHINX III software is used for measuring the recognition accuracy of the reconstructed signals. As mentioned before, the reconstruction of missing elements in the geometrical methods is based on the observed elements of the spectrogram. Two main methods in linear geometrical methods are the interpolation along time and along frequency are done in these experiments. An example of the reconstructed spectrograms obtained by the interpolation along time and frequency for a noisy spectrogram are shown in Figure 1 and Figure 2, respectively.



a b

FIGURE 1. a) An utterance spectrogram.
b) Noisy spectrogram with SNR =20 dB



a b

FIGURE 2.
a) Reconstructed spectrogram using the interpolation along frequency.
b) Reconstructed spectrogram estimating the missing using the interpolation along time.

Table 1 shows the recognition accuracies of reconstructed one-word utterance spectrogram using linear geometrical methods (interpolation along time and frequency) with different window time and window step values. While In Table 2, the recognition accuracies of reconstructed small sentence utterance spectrogram using linear geometrical methods (interpolation along time and frequency) with different window time and window step values are shown.

Sometimes there is no sufficient information in the spectrogram itself to apply the geometrical methods; this is the main drawbacks of geometrical methods. Moreover, as we know the interpolation in the geometrical methods will

be along frequency axis or time axis. In this method, the interpolation can be done in both axes according to the available observed elements. The interpolation will start for example for each spectral vector (along frequency) until we cannot make this process, the interpolation will be done in time slicing (along time). In this method, we tried to avoid the extrapolation as we can, because as we know the extrapolation is not good like interpolation for estimation. We tested this method on utterance with window time 25 ms and window step 10 ms. The results can be shown in Table 3. As shown in the Table1, the highest word recognition accuracy is achieved when window time 25 ms, and window step 10, reaching around 76% using interpolation along time and 71% using interpolation along frequency. Whereas in Table 2 the highest sentence recognition accuracies obtained using interpolation along time and interpolation along frequency are 55% and 51%, respectively. Using a combined interpolation by time and frequency, the highest word recognition accuracy is achieved, reaching 80% with SNR 25 which is outperformed the previous experiment results.

TABLE 1. Word recognition accuracy Vs. SNR values of noise where 1) Interpolation along time 2) Interpolation along frequency

Window time =40 Window step =20			**Window time =30 Window step =15**			**Window time =25 Window step =10**		
SNR	**(1)**	**(2)**	**SNR**	**(1)**	**(2)**	**SNR**	**(1)**	**(2)**
0	15	12	**0**	17	14	**0**	17	14
5	27	21	**5**	32	29	**5**	32	30
10	32	31	**10**	37	34	**10**	37	35
15	35	34	**15**	40	37	**15**	40	37
20	52	49	**20**	62	54	**20**	65	60
25	67	62	**25**	71	67	**25**	76	71

TABLE 2. Sentence recognition accuracy, Vs. SNR values of noise 1) Interpolation along time 2) Interpolation along frequency

Window time =40 Window step =20			**Window time =30 Window step =15**			**Window time =25 Window step =10**		
SNR	**(1)**	**(2)**	**SNR**	**(1)**	**(2)**	**SNR**	**(1)**	**(2)**
0	9.9	8	**0**	12	10	**0**	13	11
5	24	21	**5**	24	21	**5**	24	21
10	27	25	**10**	32	31	**10**	32	31
15	30	28	**15**	35	32	**15**	35	32
20	37	31	**20**	41	34	**20**	47	41
25	39	34	**25**	52	49	**25**	55	51

TABLE 3. Table 3 Word Recognition accuracy, Vs. SNR using new geometrical reconstruction method

Window time =25 Window step =10	
SNR	**Recognition Accuracy%**
5	30
10	35
15	45
20	60
25	80

CONCLUSIONS

Many of spectrogram reconstruction methods can be used for compensating the noise effect. The incomplete spectrogram is the result of removing all noisy regions in the spectrogram. In these methods, the missing elements (removed elements) in noisy regions are reconstructed before performing the recognition to get the complete spectrogram. One important point is that the noisy regions are identified a priori. From the complete spectrogram, the power spectral or related features can be extracted and recognition performance can be obtained. Implementing these methods as a toolbox help us to study these methods and to try to improve these methods or suggest new methods. Reconstruction of missing features helps to provide a good speech recognition accuracy results. In addition, the experiments observed that with small window size (25 ms – 40 ms) and for small utterance, good results are obtained. Some conditions like window time and window step values have an effect on the result of reconstructing the missing elements as motioned in previous section.

ACKNOWLEDGMENTS

This work is supported by the Universiti Malaysia Pahang (UMP) via Research Grant UMP RDU160349 and Research Grant UMP DRU150353.

REFERENCES

[1] M. L. Seltzer, "Automatic detection of corrupt spectrographic features for robust speech recognition." PhD thesis, Carnegie Mellon University, 2000.

[2] B. R. Ramakrishnan, "Reconstruction of incomplete spectrograms for robust speech recognition." PhD thesis, Carnegie Mellon University, 2000.

[3] Y. Wang, J. F. Gemmeke and K. Demuynck. Missing data solutions for robust speech recognition. in Essential Speech and Language Technology for Dutch (Springer, 2013), pp. 289-304.

[4] M. Cooke, P. Green, L. Josifovski and A. Vizinho, “Robust Automatic Speech Recognition with missing and unreliable acoustic data’’ Speech Communication, 34(3) P 267 - 285, (2001).

[5] B. Raj, M.L. Seltzer, & R.M. Stern, "reconstruction of missing features for robust speech recognition”, Speech Communication, 43(4), pp.275-296, (2004).

[6] X. Huang, A. Acero, Hon., "Spoken Language Processing: A Guide to Theory, Algorithm and System Development ", Prentice- Hall, Inc, 2001.

[7] P. Oliveira and L. Gomes, Incomplete data in sample surveys. Vol. 3: proceedings of the symposium. (1983).

[8] P. Oliveira, & L. Gomes. Interpolation of signals with missing data using Principal Component Analysis. Multidimensional Systems and Signal Processing, 21(1), 25, (2010).

[9] L. M. Lee, & F. R. Jean, Model adaptation method for recognition of speech with missing frames. The Journal of the Acoustical Society of America, 135(3), EL166-EL171, (2014).

[10] H. Rassem and P. Girija. Missing Features Restoration Using Clustering Methods. In the Signal-Image Technology and Internet-Based Systems (SITIS), 2010 Sixth International Conference on, 2010.

[11] P. Magron, R. Badeau and B. David. Phase reconstruction of spectrograms with linear unwrapping: application to audio signal restoration. In Signal Processing Conference (EUSIPCO), 2015 23rd European (pp. 1-5), (2015).