Comparison of interpolation methods for missing data reconstruction

Elaine Pereira Lima Scartezzini and Carlos Alberto Ynoguti

***Abstract*— The missing data approach was developed to perform automatic speech recognition in noisy environments. This technique identifies and uses in the recognition process only parts of a noisy utterance which were not heavily corrupted by the noise, these parts are called reliable. There are two main methods that can be used to achieve this goal: the marginalization and the imputation. The marginalization method uses only the utterance reliable information, whereas the imputation method tries to substitute the unreliable parts for estimates based on the reliable information. The purpose of this paper is to compare three imputation methods: the linear interpolation, the polynomial interpolation and the rational interpolation.**

***Keywords*— Speech recognition under noise, missing data, imputation reconstruction, interpolation reconstruction.**

I. INTRODUCTION

With the maturation of the speech recognition technology, it is being more and more used in several applications, such as voice dialing in smartphones, control of TVs, web browsing, etc.

On the other hand, mobile devices have to operate in very different acoustic situations. In this case, the performance of these systems dramatically degrades, mainly due to the environmental noise. Although the human ear has a huge capacity to distinguish sounds even if they are immersed in noise, this capacity is not yet fully reproduced by automatic speech recognition systems [1].

Large investments in automatic speech recognition systems are being made, mainly on mobile devices (such as mobile phones, tablets and smart watches) and appliances (as television and radio), which led to a great evolution in this technology, but its performance is not satisfactory in all environments [2].

Techniques for automatic speech recognition in the presence of additive noise have been widely studied in recent years. The main techniques are: spectral subtraction [3], cepstral mean normalization [4] and theory of missing data [5], the last one being the focus of this work.

In the missing data technique is not necessary to have knowledge about the noise and it remains robust even in high noise levels. This technique can be divided into two methods: marginalization and imputation. Both methods work with the recognition based solely on data that is not heavily corrupted by noise, but the imputation method has advantages such as: it is not necessary to modify the recognizer and it is possible to use cepstral vectors, because this method does the reconstruction of missing data.

This work presents a comparison between reconstruction methods for interpolation of missing data using the imputation, in other words, some data from a vocal composition were lost due to noise insertion and the created system will attempt to rebuild them through the interpolation of data considered reliable (no noise).

II. MISSING DATA

The missing data approach was first proposed by researchers at the University of Sheffield in the United Kingdom in the early 1990s [7]. It is based on two main ideas: a) when the speech signal is corrupted by noise, some time-frequency components are more corrupted than others; and b) the speech signal has lots of redundancies, therefore it would be possible to perform the recognition based only on high SNR components.

It is possible to estimate, before decoding, which spectro-temporal regions in the acoustic representation of a noisy speech are reliable (mainly dominated by speech energy) or unreliable (mainly dominated by background noise) by the analysis of the speech signal spectrum. A matrix called missing data mask is then created to indicate which parts of the spectrum can be considered reliable and which ones are not. There are several methods for estimating noise and identify the reliable and unreliable components, and they may be found in [12].

After defining the reliable data through the mask, two main recognition techniques can be used: marginalization and imputation [6]. In the marginalization technique, the recognition is done only with reliable data, ignoring the remaining ones. This approach requires a modification in the recognition engine in order to calculate the scores using only part of the input vectors. On the other hand, for the imputation technique, the missing data is estimated from the reliable data, allowing the recognition to be made from a full time-frequency representation, and therefore, no modification in the recognition engine is necessary. This is the technique used in the experiments of this work.

The most common methods for reconstruction of missing data in imputation are the interpolation, the correlation and the clustering [9] [10] [11]. The Interpolation method uses the reliable components closest to the missing component to reconstruct it. The Correlation uses the statistical dependencies between components in the current frame and in the neighbor frames. In the Cluster method the data are modeled as

Elaine Pereira Lima Scartezzini and Carlos Alberto Ynoguti, Instituto Nacional de Telecomunicações - Inatel, Santa Rita do Sapucaí-MG, Brazil, E-mails: elaine@mtel.inatel.br, ynoguti@inatel.br.

Gaussians mixtures and the missing data are calculated through the statistical relations inside a frame.

In this work the focus will be on the performance of different interpolation methods and their results. A brief explanation of different approaches is provided in the sequel.

A. Interpolation

The interpolation is a method that allows the values estimation of a function based on the knowledge of some samples of this function, and by assuming the function that models the data is smooth enough.

For spectrogram data it is possible to perform the interpolation on the time axis or in the frequency axis. According to [13], the interpolation over time is generally more effective than the interpolation along the frequency. Thus, in the experiments of this work, only the time axis interpolation will be considered. This interpolation is done just after the calculation of the FFT, in the mel cepstrum coefficients extraction procedure.

There are three main types of interpolation: Linear, Polynomial and Rational, which will be presented in details below.

1) Linear Interpolation: The linear interpolation uses a first-degree polynomial to represent a discontinuous function in a given range. Formally, it can be defined as follows:

Let $P_1(x)$ be a first-degree polynomial that passes through the points $A = (x_i, f_i)$ and $B = (x_{i+1}, f_{i+1})$.

Then, we have the following formula to calculate the linear interpolation:

$$f(\mu) \approx P_1(\mu) = f_i + (\mu - x_i)\frac{f_{i+1} - f_i}{x_{i+1} - x_i} \quad (1)$$

where:

• μ: is the point where it is desired to calculate the function value

• x_i: is a point for which the function value is known

• $x_{(i+1)}$: is another point for which the function value is known

• $f_{(i+1)}$: is the function value at point x_{i+1}.

• f_i: is the function value at point x_i

2) Polynomial Interpolation - Lagrange form: Let x_0, $x_1,...,x_n$, be (n + 1) distinct points and $y_i = f(x_i)$, i = 0, ...n.

$p_n(x)$ is a polynomial with degree equals or less than n, that interpolates f in x_0, ...x_n.

Compactly we can write the Lagrange form for interpolating polynomial as:

$$p_n(x) = \sum_{k=0}^{n} f(x_k) L_k(x) \quad (2)$$

where $L_k(x)$ are the Lagrange factors, given by:

$$L_k(x) = \frac{\prod_{\substack{j=0 \\ j \neq k}}^{n} (x - x_j)}{\prod_{\substack{j=0 \\ j \neq k}}^{n} (x_k - x_j)} \quad (3)$$

3) Rational Interpolation - Bulirsch-Stoer Algorithm: It is an extrapolation method that uses a rational function to approximate the solution points of an ordinary differential equation within a given range [14].

The algorithm is based on the midpoint method, a Runge-Kutta second order method. The method execution starts by applying the Euler method for obtaining a first approximation and, successively applies the midpoint method to generate subsequent approximations, performing what is known as deferred approach to the Richardson limit. Finally, it applies the extrapolation based on a rational function [15] [16]. More details about this method can be found in [17] and [18].

III. MATERIAL AND METHODS

A. Database

For the tests, the TIDIGITS database was chosen. It was originally designed and collected at Texas Instruments in 1982, with the purpose of designing and evaluating algorithms for speaker-independent speech recognition [1].

This database is composed by digits utterances from 326 speakers: 111 men, 114 women, 50 boys and 51 girls. The records are in English language and the speakers come from 21 different regions of the United States of America.

For this work the speakers are partitioned into two subsets: test and training. Although the database contains isolated and connected digits utterances, in this work, only the utterances with isolated digits were used. In this subset, each speaker pronounced 11 digits: ”zero”, ”oh”, ”one”, ”two”, ”three”, ”four”, ”five”, ”six”, ”seven”, ”eight” and ”nine”, repeated 3 times each. The data was collected in a low noise environment and digitized at 20 kHz, with 16 bits of resolution.

Only the adult speakers were used in this experiment, and the training subset consists of 57 women and 55 men, and the test subset consists of 57 women and 56 men. The test speakers are different from training speakers.

For this work, all utterances were down sampled to 8kHz through the Linux platform command “Sound eXchange” (sox) [20].

As acoustic features, the mel-cepstral coefficients, together with their first and second derivatives were chosen. These coefficients were calculated from 25 ms windows, updated at every 10 ms. Prior to parameterization, the utterances passed through a first order pre-emphasis filter with $1 - 0.97z^{-1}$ system response.

B. Recognition engine

The HTK (Hidden Markov Model) toolkit was chosen as the recognition engine. It is a portable free tool for creating and manipulating hidden Markov models, originally developed by the engineering department of the University of Cambridge (CUED).

This tool is widely used in speech recognition research, but is also used in many other applications including research in speech synthesis, character recognition and DNA sequencing [21].

C. *Methodology*

The focus of this work is the comparison of three reconstruction strategies: linear interpolation, polynomial interpolation and rational interpolation. Therefore, instead of using a missing data mask calculated from a noise corrupted utterance, it was chosen to use clean speech signals, with some spectro-temporal parts artificially removed (simulating the mask). In our understanding, this procedure leads to a fairer comparison of the imputation techniques.

Also, the chosen scenario is a speaker independent recognition of isolated digits because, in this case, the recognition has to rely solely on acoustic evidences, without the help of grammars.

The recognition system comprises 11 word models, one for each digit, each one trained with the utterances of male and female speakers from the training set.

IV. RESULTS AND DISCUSSION

A. *Baseline*

The first test is the recognition of utterances with no mask applied. In this case, an accuracy of 99.39 % for the recognition of the utterances from the testing set. This is the baseline performance for this work.

B. *Tests with imputation*

As the focus of this work is not the creation of missing data mask but it is the reconstruction of the utterance and your accuracy, we will assume that the data which will be imputed are known, that is, the mask has an ideal performance.

Interpolations are made on the time axis, where the experiments have most significant results [13]. As an example, Figure 1 represents a mel spectrogram of an utterance without noise from the database. Figure 2 shows the spectrogram points that have been imputed, and Figure 3 is a representation of this same utterance reconstructed with the linear interpolation.

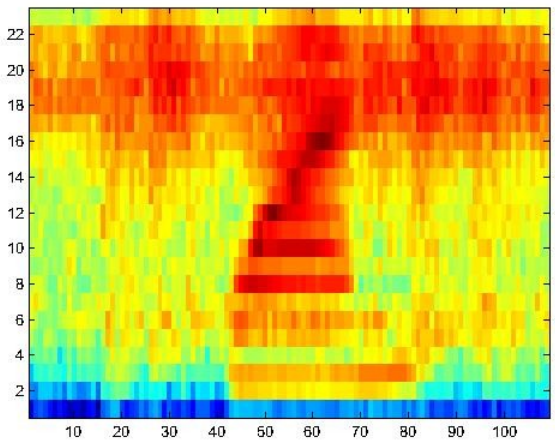

Fig. 1. Spectrogram before imputation

The comparison between Figure 1 and Figure 3 shows that it is reasonably possible to reconstruct the speech spectrogram, even when most of the original information is missing.

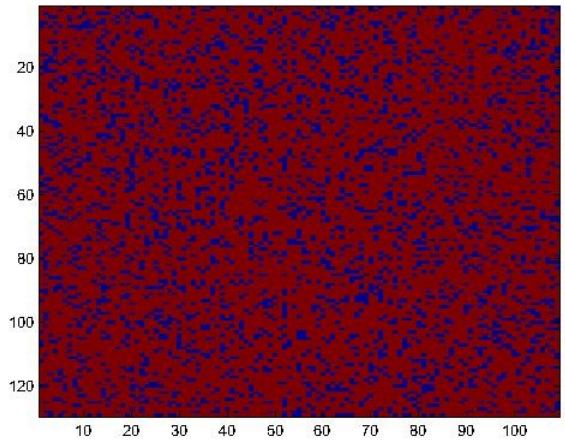

Fig. 2. Missing data mask, with 80 % of the points selected for removal (imputed points are the darker ones).

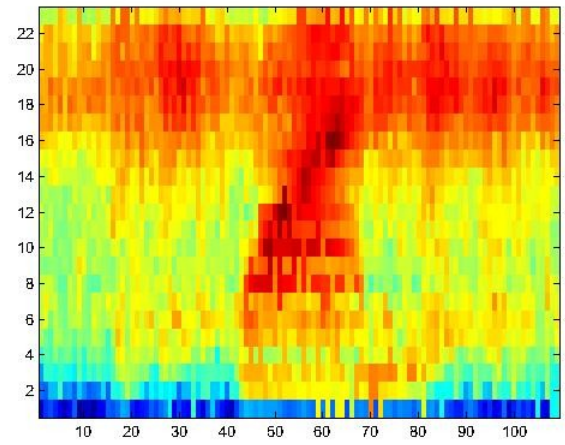

Fig. 3. Spectrogram after imputation.

The missing elements were randomly chosen, according to a Bernoulli distribution, with the probability of occurrence ranging from 10 % to 80 %, in steps of 10 %. The results obtained for each method are shown in Figure 4.

All three methods have similar performance, with little advantage for the linear interpolation. Also, it is possible to see that the imputation methods work well even when very little amount original information is available: the performance remains steady until the 50% - 60% range, dropping dramatically only at the 70% region.

V. CONCLUSIONS AND FUTURE WORK

In this work, three imputation methods for the missing data approach were compared: the linear interpolation, the polynomial interpolation and the rational interpolation, for speaker independent, isolated digit recognition task.

All three methods performed in a similar way, with little advantage for the linear interpolation, which is a double gain: better performance combined with lower complexity. The

recognition rates stayed close to the baseline performance even with 50 % - 60 % of information missing.

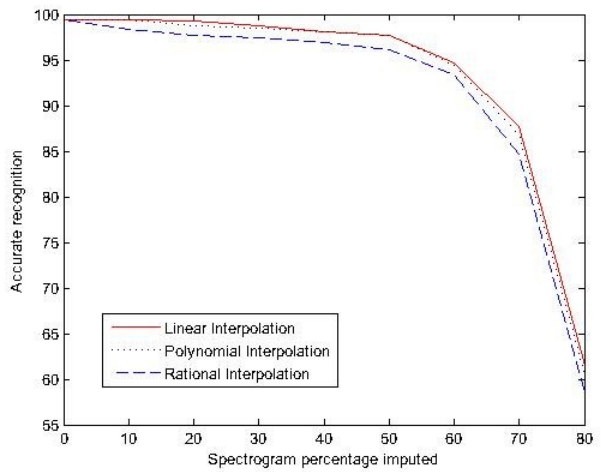

Fig. 4. Results

For the future, the use of non ideal masks, calculated from the actual noisy speech has to be tested, in order to verify the real performance of this idea.

REFERENCES

[1] Richard P. Lippmann, Speech recognition by machines and humans, Lincoln Laboratory MIT, April 1997.

[2] M. Anusya and S. Kati, Speech Recognition by Machine: A Review, International Journal of Computer Science and Information Security, Vol. 6, 2009.

[3] A. R. Fukane, S. L. Sahare, Different Approaches of Spectral Subtraction method for Enhancing the Speech Signal in Noisy Environments, International Journal of Scientific & Engineering Research, vol. 2, May2011.

[4] F. Liu, R. M. Stern, X. Huang and A. Acero, Efficient cepstral normalization for robust speech recognition, Proceedings of ARPA Speech and Natural Language Workshop, 1993.

[5] B. Raj and R.M. Stern, Missing-feature approaches in speech recognition, Signal Processing Magazine, IEEE , vol. 22, pp.101-116, Sept.2005.

[6] B.Raj and R. Singh, Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications, chapter 6, pp. 127-156, New York, 2011.

[7] Dorothea Kolossa and Reinhold Haeb-Umbach. Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications, Springer Verlag, 380 pages, July 2011.

[8] S. Hsiang, Missing-Feature Approaches in Speech Recognition, SLP2006.

[9] Jort Florent Gemmeke and Ulpu Remes. Missing data techniques: Feature reconstruction, Journal Techniques for Noise Robustness in Automatic Speech Recognition, 2012.

[10] J. F. Gemmeke and U. Remes, Missing-Data Techniques: Feature Reconstruction in Techniques for Noise Robustness in Automatic Speech Recognition, chapter 15, October 2012.

[11] K. Wagstaf , Clustering with Missing Values: No Imputation Required in Studies in Classification, Data Analysis, and Knowledge Organization, pp. 649-658, July 2004.

[12] Cerisara, C., Demange, S., Haton, J.-P., On noise masking for automatic missing data speech recognition: a survey and discussion, Computer Speech and Language, 21(3):443-457, July 2007.

[13] Bhiksha Raj Ramakrishnan, Reconstruction of Incomplete Spectrograms for Robust Speech Recognition, Department of Electrical and Computer Engineering Carnegie Mellon University, 2000.

[14] Thiago Alves de Queiroz e Donal Mark Santee, Um aprimoramento no método de extrapolação de Gragg-Bulirsch-Stoer para obter alta precisão e baixo esforço computacional, 2007.

[15] R. Bulirsch and J. Stoer, Numerical Treatment of Ordinary Differential Equations by Extrapolation Methods, Numerische Mathematik 8, 1-13 (1966).

[16] S. Kirpekar, Implementation of Bulirsch Stoer Extrapolation Method, Department of Mechanical Engineering, UC, Berkeley/California (2003).

[17] W. B. Gragg, On extrapolation algorithms for ordinary initial value problems, SIAM J. Num. Anal. 3, 384-403 (1965).

[18] N. L. Schryer, An Extrapolation Step-Size Monitor for Solving Ordinary Differential Equations, Proceedings of the 1974 annual conference (ACM/CSC-ER), 140-148 (1974).

[19] TIDIGITS, Joseph Picone. Available at: http://www.isip.piconepress.com/projects/speech/software/tutorials/production/fundamentals/v1.0/section 02/s02 04 p01.html (last accession March 2015)

[20] SoX - Sound eXchange. Available at: http://sox.sourceforge.net (last access: on March 2015).

[21] Hidden Markov Model Toolkit (HTK) Speech Recognition Toolkit. Available at: http://htk.eng.cam.ac.uk/ (last access: on April 2015).