International Journal of Scientific Research and Engineering Development-– Volume 2 Issue 5, Sep – Oct 2019

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Noise robust digit recognition using sparse representations

Ankit Narendrakumar Soni*

*Department of Information Technology, Campbellsville University

Email: soniankit.ra@gmail.com

----------**********----------

Abstract:

Despite the utilization of noise hardiness techniques, automatic speech recognition (ASR) systems create more recognitionerrors than humans, particularly in terribly creaking circumstances. We argue that this inferior recognition performance is essentially due to the actual fact that in ASR speech is usually processed on a frameby-frame basis preventing the redundancy within the speech signal to be optimally exploited. we tend to gift a completely unique non-parametric classification methodology that may handle missing information whereas at the same time exploiting the dependencies between the reliable options in a whole word. we tend to compare the new methodology with a state-of-the-art HMM-based speech decoder within which missing data area unit imputed on a frame-by-frame basis. each ways area unit tested on one digit recognition task (based on AURORA-2 data) exploitation associate degree oracle associate degreed an calculable harmonicity mask. We show that at associate degree SNR of -5 decibel exploitation the reliable options of associate degree entire word permits associate degree accuracy of ninety one (using mel-log-energy features together with associate degree oracle mask), whereas a standard frame-based approach achieves solely sixty one. Results obtained with the harmonicity mask recommend that this specific mask estimation technique is just unable to deliver decent reliable features for acceptable recognition rates at these low SNRs. Index Terms: Non-parametric model, Compressed Sensing, Missing information Techniques

Keywords **—Automatic speech recognition, exemplar-based, noise robustness, reconstruction error, sparse representations.**

----------**********----------

I. INTRODUCTION

Humans exceed automatic speech recognition (ASR) systems, especially once the speech is severely degraded by noise. While a lot of progress has been created within the past few decades, at

low SNRs, ASR is therefore very little noise strong that it becomes just about useless. during this paper we have a tendency to argue that this inferior performance at low SNRs is essentially because of the frame-based approach that's typically utilized in progressive ASR systems. Missing knowledge Techniques (MDT) [1, 2] have tested a robust means to mitigate the impact of noise on recognition accuracy.

The general plan behindMDT is that it's doable to estimate−prior to decoding− that spectro-temporal parts of the acoustic representations area unit reliable (i.e., dominated by speech energy) and that area unit unreliable (i.e., dominated by background noise). By storing these reliableness estimates in an exceedinglyso referred to as prism spectroscope mask, this data are often wont to treat reliable options otherwise from unreliable ones throughout decoding: One might either impute the unreliable options [3, 4], or one might use a social process approach within the decoder when evaluating unreliable input file [2]. The continuity constraints obligatory upon the speech signal by the speaking system area unit thought-about to represent associate degreeimportant supply of redundancy. In

most ASR approaches, but but, remarkably very little effort is spent on exploiting this data. Speech is generally processed on a frame by frame (i.e. strictly local) basis. The disadvantages of this approach become particularly evident at low SNRs (⩽ zero dB). In those conditions,it may happen that solely few, if any, parts in associate degree acoustic vector are labelled reliable. and clearly, the less reliable options remain, the a lot of serious the chance that a private frame contains deficient data for correctly coping with the unreliable coefficients. It thus looks logical to presume that, if too several frames with solely some reliable options exist, recognition accuracy can suffer considerably which this downside can solely be relaxed by creating higher use of the reliable options in neighbouring frames likewise. In this paper we have a tendency to gift associate degree innovative classification technique recently introduced within the field of face recognition [4] that allows United States of America to review the importance of wider time context.This technique, hereafter noted as distributed classification, is an application of compressed sensingand performs classification by wanting however well associate degree determined sequence of feature vectors are often explained by a linear combination of example speech sequences from an equivalent category. This non-parametric approach requires no coaching and is quickly extended to handle missing knowledge since the spatiality of the determined speech signal doesn't ought to be fastened earlier as critical once using a constant quantity model.

II. PROPOSEDMETHOD

A. Speech data and classification task

The classification technique delineated within the following sections works on observation vectors of mounted size. This makes it not directly applicable to running speech. during this paper, we have a tendency to so restrict ourselves to a single-digit recognition/classification task. The single-digit utterances were created by repetition all individual digits from the AURORA-2 corpus. The section boundaries were obtained via a forced alignment of the clean speech utterances with the reference transcriptions. We used only check set A, that includes one clean and twenty four rip-roaring subsets, with four noise sorts (subway, car, babble, exhibition hall) at six SNR values, SNR= 20, 15, 10, 5, 0,−5 dB to guage recognition accuracy as a operate of classification technique and SNR.

B. Baseline speech decoder

As our baseline system, we have a tendency to used a MATLAB implementation of a missing information recognition system delineate in [4]. Acoustic feature vectors consisted of sixty nine PROSPECT options,created from twenty three mel frequency log power spectra additionally as their initial and second derivatives. options that area unit labelled as unreliable (by some outwardly provided spectroscope mask) are replaced by calculable values victimization most chance per Gaussian-based imputation. we have a tendency to trained eleven whole-word models with sixteen states per word, additionally as 2 silence words with one and three states severally, victimization clean speech.

C. Fixed length vector representation of digits

To obtain a set length feature vector for every digit, pro re nata by the thin representation method in section a pair of.4, we have a tendency to regenerate the variable variety of acoustic vectors creating up a word unit (originally at a set frame rate of one hundred Hz) to a time normalized version (with a set variety of acoustic vectors at a variable frame rate). A spline interpolation was applied to the individual time tracks of all mel frequency log-energy coefficients once which they were re-sampled so thirty five acoustic vectors per digit resulted (i.e., the mean variety of ten ms time frames per word in the coaching set). Next, these time-frames were concatenated to form one, mounted length observation vector y per digit with a dimension K = twenty three · thirty five = 805.

D. Sparse representation

Following [5] we tend to think about a take a look at digit y to be a linear combination of ideal digits di,n, wherever the primary index (1 ≤i ≤

I)denotes one among the I = eleven digitcategories and therefore the second index (1 ≤ n ≤ Ni) a selected idealdigit of sophistication i with nickel the quantity of ideal digits in every class. We write:

$$y=\sum_{i=1}^{I}\sum_{n=1}^{N_i} \propto i,n,d_i,n$$

E. l^l minimization

In order to represent a digit y by the thin vector x one desires to solve the system of linear equations of relative atomic mass. 1. Typically, the number of ideal digits are a lot of larger than the spatial property of the feature illustration of the vowels (N ≫ K).Thus, the system of linear equations in relative atomic mass. one is underdetermined and has no distinctive answer. analysis within the field of compressed sensing [6, 7] has shown that if x is thin, x may be recovered by solving:

$$\min\|x\|1 \text{ subject to } y = Ax \qquad (2)$$

with ‖.‖1 the l1norm (i.e. minimisation of the total of absolute values of elements) that is associate approximation of the l0 norm (i.e., the quantity of nonzero elements). The approximation is necessary since minimizing the l0 norm is combinatorial problem is associate NP-hard [9] whereas l1minimization may be done efficiently in polynomial time. Since in apply it's going to be not possible to express a digit specifically as a superposition of ideal digits, we have a tendency to use a noise sturdy version of relative atomic mass. 2 (cf. [10]):

$$\min\|x\|1 \text{ subject to } \|y - Ax\|2 \leq \in \qquad (3)$$

withatiny low atiny low such the error e satisfies ‖e‖2 <∈.

III. Discussion

The recognition accuracy of ninety one at SNR = -5 dB obtained with the SC-method victimization Associate in Nursing oracle mask shows that at terribly low SNRs enough info concerning the speech signal is preserved to with success perform classification alone on the premise of reliable time-frequency cells, even once the acoustic illustration consists of classical frequency filterbands. Comparing this to the sixty one recognition accuracy of the baseline decoder, it is clear that the data contained in reliable options is only partly used once doing missing information recognition on a frame-by-frame basis. The drop by accuracy at all-time low SNR's from ninety nine to ninety one is mainly thanks to digits that have only a few or maybe no reliable features. Apparently, it often happens that merely insufficient reliable information is left to base recognition on. This result is not sudden and an identical drop by recognition accuracy can be discovered for human subjects at negative SNRs. Using Associate in Nursing oracle mask, SC performs slightly worse than the NC of the baseline recognizer for SNR ⩾ five dB.this means that the SC methodology doesn't generalize to discovered digits as well because the HMM-based approach. There is also many reasons for this. Possibly, the premise size is just too tiny and therefore the accuracy gap might be closed by employing a larger basis. Also, the dimensionality reduction given in Sec. 2.8, whereas greatly reducing the machine price, may need a small, adverse effect on recognition accuracy. Finally, the baseline recognizer uses MFCC-like (i.e. PROSPECT) options whereas our classification method works directly on mel log-energy coefficients. Future analysis is required to what extent these factors play a significant role. Using the SC methodology with Associate in Nursing calculable prism spectroscope mask, i.c. the harmonicity mask, we tend to acquire recognition accuracies that ar well under those with Associate in Nursing oracle mask. The fact, that the SC methodology performs even worse than the American state of the baseline recognizer will solely mean that the number of reliable features known in and of itself by the harmonicity mask is just too small to warrant a correct recognition. Comparing the accuracies obtained with SC together with the oracle mask on the one hand and people obtained with NC together with the harmonicity mask on the opposite (cf. Fig. 1), one might observe that the values of the primary curve

attain approximately identical values because the latter, however at SNRs that are ten dB higher Fig. a pair of reveals that an identical relation holds for the percentage of reliable cells (the underdeterminedness): The percentage of reliable cells found with the harmonicity mask is roughly identical like Associate in Nursing oracle mask at noise levels that are ten dB apart. forward that the fraction of reliable options per word is determinant the utmost accomplishable recognition accuracy, this means that the baseline recognizer already has reached a ceiling which the low accuracies at lower SNRs should be attributed to the actual fact that there ar merely not enough reliable coefficients left.

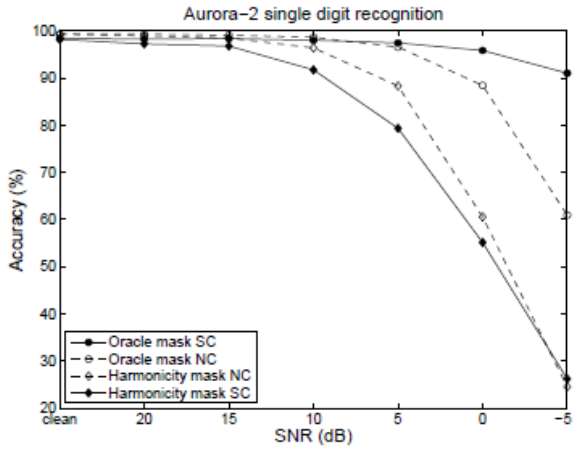

Fig. 1 AURORA-2 single digit recognition accuracy. The figure shows results for both normal classification (NC) and sparse classification (SC) for the oracle mask and the harmonicity mask

A attainable clarification for the very fact that the SC methodology performs worse than Tar Heel State is that the baseline decoder uses info not on the market to the SC methodology. In distinction to the SC methodology which makes no assumptions regarding the unreliable options, the baseline recognizer assumes that the energy of the speech signal in these time-frequency cells cannot exceed the ascertained energy. Without these bounding constraints, the SC methodology would possibly also be additional sensitive to false reliables that, as shown in Fig. 2, conjointly become additional various as noise levels increase. While not applied within the current study, the distributed classification methodology could simply be extended with an identical constraint.

IV. CONCLUSIONS

We introduced a non-parametric missing information classification method that works by finding a thin illustration of the noisy speech signal, exploitation solely the reliable info of the speech signal as tagged by a prism spectroscope mask. The method exploits the redundancy of the speech signal within the time frequency domain by expressing entire words as a linear combination of model speech signals. wetend to showed the potential of the method by achieving recognition accuracies on AURORA-2 digits of ninety one at SNR -5 sound unit exploitation AN oracle mask, a rise of 30% p.c absolute over a state-of-the art missing information speech recognizer. These findings show that a lot of progress will still be made exploitation typical options by adapting the coding algorithms so that the redundancy within the time domain is correctly exploited. However, the popularity accuracies obtained with the harmonicity mask additionally indicate that this mask estimation technique would possibly merely deliver too few reliable time-frequency cells to alter really noise sturdy recognition.so as to be able to very make the most of these insights future work can want to target techniques that exploit the redundancy properties of speech already throughout the mask estimation procedure.

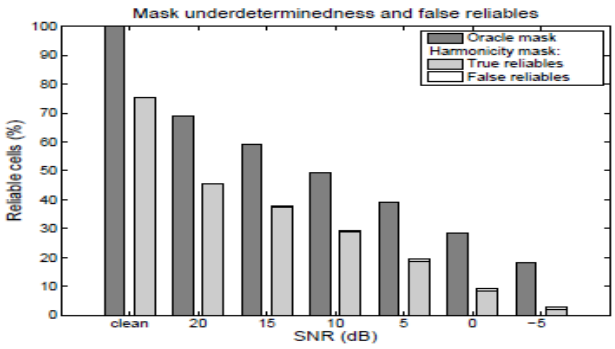

Fig. 2 percentage of reliable time-frequency cells for the oracle and to the oracle mask, are falsely labeled reliable.

REFERENCES

[1]. B. Raj, R. Singh, and R. Stern, "Inference of missing spectrographicfeatures for robust automatic speech recognition," in ProceedingsInternational Conference on Spoken Language Processing, 1998, pp. 1491–1494.
[2]. M. Cooke, P. Green, L. Josifovksi, and A. Vizinho, "Robust automaticspeech recognition with missing and unreliable acoustic data," Speech Communication, vol. 34, pp. 267–285, 2001.
[3]. B. Raj, "Reconstruction of incomplete spectrograms for robustspeech recognition," Ph.D. dissertation, Camegie Mellon University, 2000.
[4]. H. Van hamme, "Prospect features and their application tomissing data techniques for robust speech recognition," in INTERSPEECH-2004, 2004, pp. 101–104.
[5]. Vishal DineshkumarSoni. (2018). Artificial Cognition for Human-robot Interaction. International Journal on Integrated Education, 1(1), 49-53. https://doi.org/10.31149/ijie.v1i1.482 9.
[6]. Vishal DineshkumarSoni. (2018). An IoT Based Patient Health Monitoring System. International Journal on Integrated Education, 1(1), 43-48. https://doi.org/10.31149/ijie.v1i1.481
[7]. D. L. Donoho, "Compressed sensing," IEEE Transactions on Information Theory, vol. 52, no. 4, pp. 1289–1306, 2006.
[8]. E. J. Candes, "Compressive sampling," in Proceedings of the International Congress of Mathematicians, 2006.
[9]. 'An Advanced Algorithm for Cancer Detection Using Image Processing Techniques' International Journal of Science and Research (IJSR), volume 4, issue 4, April-2015, ISSN 2319-7064, IMPACT FACTOR 6.2
[10]. 'Modulated Backscattering Coverage in Wireless Passive Sensor Networks' International Journal of Science and Research (IJSR), volume 3, issue 1, Jan-2014. ISSN 2319-7064. IMPACT FACTOR 6.2
[11]. 'An Efficient Architecture for Lifting Based 3D-Discrete Wavelet Transform' International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 12, December – 2013 ISSN: 2278-018

