Evaluating the WER for different Extraction Methods

Saimir Tola[1], Alfred Daci[1]
Polytechnic University of Tirana Faculty of Mathematical and Physical Engineering, Tirane, Albania,[1]
Polytechnic University of Tirana Faculty of Mathematical and Physical Engineering, Tirane, Albania,[1]
saimir_tola@yahoo.com

Abstract: *In this paper we will discuss an important topic such as the WER (word error rate). Imagine if we use an ASR system in a real event for approximately one hour or more. We would have a lot of issues like: the quality of the transcription of words, the time of the processing of the words spoken. And so a lot of WER will be produced by this event. The high rate of ASR errors have demanded the necessity to find better techniques in order to correct such errors. The improvement of the system it is a necessity not only to have a better stability of the system but also to prevent the high costs in order to use our resources in the best way as possible.*
KEYWORDS*: WORD ERROR RATE, AUTOMATIC SPEECH RECOGNITION*

1. Introduction

Automatic Speech Recognition systems is used in order to converting a speech signal into a sequence of words. The purpose of evaluating ASR systems is to improve or to make it easy for the humans in order to measure their utility especially when comparing systems. The standard metric of ASR evaluation is the Word Error Rate, which is defined as the proportion of word errors to words processed. ASR has matured to the point of commercial applications by providing transcription with an acceptable level of performance which allows integration into many applications. In general, ASR systems are effective when the conditions are well controlled. Nevertheless, they are too dependent on the task being performed and the results are far from ideal, and especially for Large Vocabulary Continuous Speech Recognition (LVCSR) applications. This later still one of the most challenging tasks in the field, due to a number of factors, including poor articulation, variable

Speaking rate and high degree of acoustic variability caused by noise, side-speech, accents, sloppy pronunciation, hesitation, repetition, interruptions and channel mismatch, and/or distortions. To deal with all these problems, there has been a plethora of algorithms and technologies proposed by the scientific communities for all steps of LVCSR over the last decade: pre-processing, feature extraction, acoustic modeling, language modeling, decoding and result post-processing. Nevertheless LVCSR systems are not yet robust with error rates of up to 50% under certain conditions [1],[3].The persistent presence of ASR errors motivates the attempt to find alternative techniques to assist users in correcting the transcription errors or to totally automate the correction process. Manual errors correction is often tedious and time consuming. Hence automatic detection and correction of ASR errors has become an important research area, not only for improving speech recognition accuracy but also for avoiding the propagation of the errors to the post recognition process (e.g. Machine translation and Human-Computer interaction). The aim is to be able to automatically detect, classify, and then partially or fully correct errors, regardless of the ASR system used. This can be very effective, and particularly when the ASR system is used as a black-box and the user does not have access to tune the features, the models or the decoder of the ASR system. In the present paper we present an overview about ASR errors and the stat-of-the-art techniques for their detection and correction so as to provide a technological perspective and an appreciation of the fundamental progress that has been made in this field [1].

The performance of any ASR system is evaluated in function of the error rate. The aim of ASR evaluation isto provide a comparison criterion between different systems or techniques and to measure the performance and the progress on specific tasks based on errors statistics. There are two key areas related to ASR errors, the first one is the reference-recognized alignment which consist of finding the best word alignment between the reference and the automatic transcription and the second one is the evaluation metrics measuring the performance of the ASR systems.

2. Wavelet transform (wd)

The Wavelet Transform is considered to be suited for speech processing because of its similarity to how the human ear processes sound. It is a multi-evolutional and multi-scale analysis. The three methods of wavelet transform utilized are Discrete Wavelet Transform (DWT), Wavelet Pattern Decomposition (WPD) and Discrete Wavelet Pattern Decomposition (DWPD) are given below. Information about non-stationary signals like audio can be extracted by using DWT as it is a relatively recent and computationally efficient technique for feature extraction. WPD is simply a generalization of DWT and it is a more flexible and detailed method than DWT. In WPD, the speech signal is decomposed into low frequency components and high frequency components at each level like in DWT but the key difference between DWT and WPD is that the discrete wavelet transform is introduced to the low pass result. WPD differs as it applies the transform step to both the low pass and the high pass result. The main advantage of a DWPD algorithm is that it decomposes both high frequency bands into more partitions but additionally saves complexities in computation. Wavelet Transform techniques have been shown to improve the efficiency of ASR system [2].

2.2 Improvements in the existing ASR system

It is necessary to develop unique hybrid methods that will lead to high performing ASR applications. In order to obtain better accuracy, in prosodic, text pre-processing and pronunciation fields there is still a lot of research and innovations that are needed. Mel filters; Linear Prediction (LP) and Gamma tone filters have proven to be effective features for speech and speaker recognition tasks. MFCCs as was previously discussed are usually computed by integrating short-term spectral power using a Mel-scaled filter bank (Mel FB) that typically consists of overlapping triangular filters with GFCC being the Gammatone filter equivalent. Both MFCC and PLP features perform well under matched training and test conditions but the performance gap between automatic speech recognizers (ASRs) and human listeners in real world settings is significant [3].

Varied operating conditions during signal acquisition such as channel response, handset type, ambient background noise, reverberation and so on leads to features not being correctly matched across training and test utterances, thereby degrading the performance of the MFCC. LP, and GFCC based recognizers.

In [2] et al, it is stated that warping the DFT directly instead of using filter bank averaging provides “a more precise approximation of the perceptual scales”. This was a study on additive noise degradation in ASR systems. There is a large body of research on improving the robustness of speech recognition systems under adverse acoustic environments. Environment compensation methods can be applied at the front end feature domain or at the back end model domain or can be applied on areas of the ASR system. Fast Fourier Transform (FFT) and Discrete Fourier Transform (DFT) are fundamentally alike; with the only difference being that FFT is faster. Warped DFT or FFT based features have

been found to provide lower recognition error rates than the DFT based cepstral features.

In the conventional MFCC front-end, processing of a speech signal begins with the pre-processing stage. This involves DC removal and pre-emphasis using a first-order high-pass filter with a transfer function followed by a Fourier transform being applied as was previously discussed. Transforming a linear frequency scale to a non-linear frequency scale is called frequency warping. One technique to achieve frequency warping is to apply a nonlinearly-scaled filterbank, such as a mel filterbank, to the linear frequency representation. Another way is to use a conformal mapping, such as the bilinear transformation which preserves the unit circle .

$$H(z)=\frac{z^{-1}-\alpha'}{1-\alpha' z^{-1}},\ \forall\ -1<\alpha'<1$$

The equation above is an example of warped DFT. DFT is achieved by applying the FFT algorithm. In warped DFT or FFT the positions of the frequency peaks are modified by using an all-pass transformation to warp the frequency axis. Then, uniformly-spaced points on the warped frequency axis are similar to non-uniformly spaced peaks on the original frequency axis. By picking the warping parameters sensibly, one can place some of the frequency samples in close proximity to each other to provide higher resolution in the frequency range of interest without increasing the length of the DFT. Utilizing this frequency warping, one can improve the spectral representation of speech signals in the low frequency region [4].

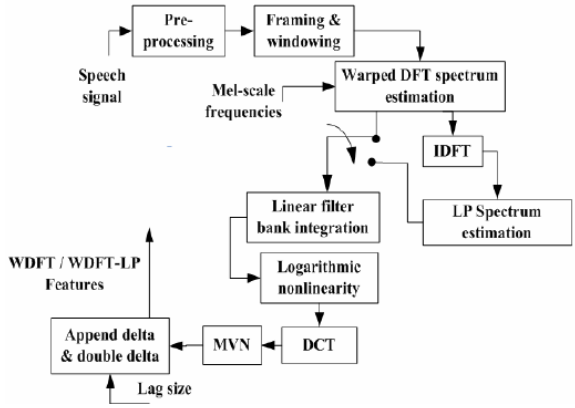

Figure 1: *Extraction of warped DFT-based cepstral features*

As stated by, warping the DFT spectrum directly without using filter bank averaging provides a more precise approximation of the perceptual scales [4]. Since the spectrum is already pre-warped using Mel-frequency warping, the nonlinearly-spaced triangular-shaped Mel-frequency filter bank is replaced by a filter bank of uniformly spaced, half-overlapping triangular filters, to provide spectral smoothing. The figure below shows running speech spectra of (a) clean and (b) noisy speech signals that have been corrupted by babble noise with a signal-to-noise ratio of 6 dB, obtained using DFT, WDFT (Warped DFT), and WDFT-LP spectrum estimators. Based on this visual examination, WDFT and WDFT-LP provide more robust spectral estimates compared to DFT and LP methods. Due to reduced degrees of freedom in all-pole modeling (model order p = 24 coefficients versus N = 256 bins), the WDFT-LP spectra are generally much smoother than the WDFT. This potentially results in improved noise robustness over WDFT [4].

Variants of the Mel-frequency warped DFT (FFT) was found to be a more robust warped frequency representation based cepstral feature, are presented in Table 1 and 2 below. From the data in Table 1 it can be seen that the WDFT based cepstral features did better on average, the DFT based MFCC and PLP features. Comparing the results of Tables 1 and 2 it can be clearly adduced that WDFT-based cepstral features performed better than the MFCC and PLP both in clean and multi-condition training modes. MFCC features computed from the Mel-warped DFT spectrum-based front-ends (WDFT, WDFT-LP) provided lower recognition error rates than the conventional MFCC and PLP on the AURORA-4 corpus. The presented speech spectra and experimental speech recognition results on the AURORA-4 LVCSR task demonstrated the robustness of the WDFT and WDFT-LP based cepstral features [3].

clean	A	B	Avg.
LP	10.31	48.78	29.545
GAMMATONE	10.79	41.45	26.12

Table 1: *Word error rates (WERs in %) obtained by the various feature extractors considered in this paper, on the AURORA-4 LVCSR corpus under clean training conditions. The model order selected in this task is: p = 24 for WDFT-LP and p = 14 for PLP. The lower the WER the better is the performance of the feature extractor.*

4. Conclusions

The experimental results above indicates that warping the DFT or FFT spectrum directly provides a more precise approximation to the perceptual scales than using filter bank averaging. It is also important to note that by applying the DFT spectrum directly, avoids using the filter banks in the conventional manner. Additionally, as evidenced by the primary research conducted , it has been shown that directly using FFT spectral values using features that that incorporate a PLP scale; can bypass the filter bank step in speech processing.

5. References

[1] G. Aist, J. Allen, E. Campana, C. Gallo, S. Stoness, Mary Swift, and Michael K. Tanenhaus. 2007. Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods. In Proc. DECALOG, pages 149–154.

[2] D. Schlangen, T. Baumann, and M. Atterer. 2009. Incremental reference resolution: The task, metrics for evaluation, and a Bayesian filtering model that is sensitive to disfluencies. In Proc. SIGdial, pages 30–37..

[3] J.D. Williams and S. Balakrishnan. 2009. Estimating probability of correctness for ASR N-Best lists. In Proc. of SIGdial 2009, pages 132–135.

[4] Alam, M. J., Kenny, P. Dumouchel, P., O'Shaughnessy, D.