Analysis of Detected Silent Segments in Call Center Recordings

Şükrü Ozan
Leonardo O. Iheme
AdresGezgini Inc.
R&D Department
İzmir, Türkiye
sukruozan@adresgezgini.com, leonardoiheme@adresgezgini.com

***Abstract*—Interpreting speech signals by making a speech to text translation is an active research area especially in current machine learning/deep learning literature. The speech to text translation of call center recordings is an important and specialized application for speech to text translation. Detecting silence in audio recordings can be a pre-processing step in order to optimize processing speed by not-considering audio parts not having significant information. In this work, such a pre-processing framework for detecting silence parts in an audio signal is considered. It is shown that further statistical analysis on the silence distributions results in detecting interesting audio features which can help in finding audio recordings which do not have actual speech sound but a fax machine tone sequence. This foundation can be directly implemented in a call center management software and makes it possible to discriminate between a normal conversation recording and a fax sound recording.**

***Index Terms*—pre-processing audio signal, statistical audio analysis, variance, entropy.**

I. Introduction

Using speech to text translation in order to interpret an audio recording became a common application after recent advances in machine learning literature together with the advance in hardware technology which is directly related with the computation power.

For a convenient speech to text translation, detection of silence in an audio sequence is an important pre-processing step since detecting and disregarding silence directly reduces the computation time [1]. After successfully removing silence segments, an audio signal can be processed by using a suitable end-to-end framework such as a recurrent neural network (RNN) as described in [2]. Silence detection is an active research area with several implementations. There are rule-based methods for silence detection, such as Energy thresholding, which have yielded good results [3]. The advantage of rule-based methods is that they do not require training however, they lack robustness since the thresholds are static and determined empirically.

In this study, we investigate the pattern of silent segments in telephone calls. We believe that this will lead to detection of calls that do not contain any speech information. Accordingly, the contribution of this study is the analysis of the periodicity of silent segments in telephone conversations for the detection of fax tones.

This work is a part of TÜBİTAK TEYDEB 1507 project with project number 7170694.

Fig. 1. Depiction of a typical non-silence/silence sequence.

The paper is organized as follows: the methods applied and the data used in the experiments in the study are explained in Section II. In the next Section we briefly outline the experimental set-up before providing the results and insights in Section IV. Finally, conclusions are drawn in Section V.

II. Methods

A. Detection of Silence

We detect silence by computing an adaptive amplitude (loudness) threshold, Th, for short segments, S, of audio files. Concretely, for each normalized audio segment S_i, we consider signals that are less than twice the Decibels relative to full scale (dBFS) as silent. The threshold was calculated using the formula in Equation 1.

$$Th_{S_i}^{[i]} = -2 * \left[20log10 \frac{||S_i||}{max(S_i)} \right] \tag{1}$$

B. Feature and Metric Selection

Any given audio sequence can be represented as consecutive non-silence/silence pairs as depicted in Fig. 1. For a typical customer-call center agent conversation, adjacent $l(n_i)$ and $l(s_i)$ values differ. Moreover, most of such typical conversations do not contain a significant amount of silence segments. However, if multiple silence segments are detected, the statistical analysis of these segments can lead to the revelation of inconspicuous characteristics in an audio recording.

where

- n_i : i^{th} non-silence segment

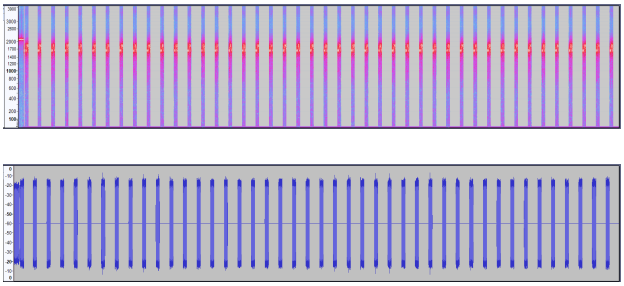

Fig. 2. A sample spectrogram and wave plot of a fax tone.

- s_i : i^{th} silence segment
- $l(x)$: length of a segment
- N : The maximum value of i in an audio sequence

It has been empirically found that the above definitions can, if used appropriately, capture periodicity in a call recording. This includes especially the characteristic periodic tones associated with fax lines when they care called.

Detecting these sounds is beneficial to the performance of a call center, since call center agents tend to perform fraudulent practices such as calling fax numbers or telephone exchange systems. In both cases, the characteristics of the signals exhibit a high degree of periodicity as is shown in the spectrogram and wave plot of Fig. 2.

The selected features are:

- $l(n_i)$: length of i^{th} non-silence segment,
- $l(s_i)$: length of i^{th} silence segment, and
- $r(i)$: $l(s_i)/l(n_i)$ ratio of the i^{th} silence segment to the adjacent non-silence segment

To find the degree of periodicity, we explore two statistics, namely:

- $\sigma^2(x)$: variance of variable x
- $H(x)$: Entropy of variable x

The formal definitions of variance (see Equation 2) and entropy (see Equation 3) suffice for this study without the loss of generality. In these equations, for an input sequence with N elements, x, $\bar{x}$ is the mean value and $p(x_i)$ represents the probability of the i^{th} sample.

$$\sigma^2(x) = \frac{\sum_{i=0}^{N-1}(x_i - \bar{x})^2}{N-1} \tag{2}$$

$$H(x) = -\sum_{i=0}^{N-1} p(x_i)\, log(p(x_i)) \tag{3}$$

The respective distributions of the six features, in addition to the combined silence and non-silence segments is depicted in Fig. 3

C. Training Data and Performance Metric

The most important prerequisite for succeeding in solving any machine learning problem is the quality of the data. In the specific case of supervised machine learning, the training data must be carefully labelled even though this could be a time consuming process.

For this study, our data set consisted of 10000 call center recordings, 1% of which were pure fax tones (positive samples) and the rest were regular call center conversations (negative samples). 20% of the data set was used as the test set and the remaining 80% was used as the training set.

In order to measure the performance of a classifier, different measures can be used. Precision is the proportion of correct positive identifications (true positives TP) to overall positive identifications (true positives + false positives TP+FP). Recall is the proportion of positive identifications to actual positives (true positives + false negatives TP+FN) [4].

When the training data is unbalanced, using either precision or recall as a model evaluation metrics is not appropriate. Instead, the F1 score [5], which takes both precision and recall into account, has been found to be more convenient. It is defined as the harmonic mean of precision and recall (see Equations 4-6) and is between 0 and 1 where 0 is the worst F1 score and 1 is the best.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{5}$$

$$\text{F1} = 2\frac{Precision * Recall}{Precision + Recall} \tag{6}$$

In this study, we sought to find the entropy and variance thresholds which maximize the F1 score. Keeping the number of negative samples 100 times more than positive samples is due to the fact that the frequency of occurrence of the fax calls in call centers is quite low. In order to mimic this nature it is more convenient to constitute this unbalanced data distribution.

III. Experimental Set-up

The features and metrics combination yielded eight different experiments for parameter estimation. The optimization criteria for parameter estimation is the F1 score which is commonly used in machine learning when there is an imbalance in the data set.

From the definition of variance, we can infer that a highly periodic distribution such as fax tones, should yield a low variance (close to zero). More precisely, the variance, σ^2 is inversely proportional to the periodicity, P i.e. $\sigma^2 \propto 1/P$. Since entropy measures the amount of "disorder" of a system, we expect that the entropy, like the variance, will be inversely proportional to the periodicity [6], i.e. $H \propto 1/P$.

In essence, we expected to get low variance and entropy values for an audio signal having a periodical distribution of silence compared to an ordinary audio file since it has a more irregular distribution of silence and non-silence segments.

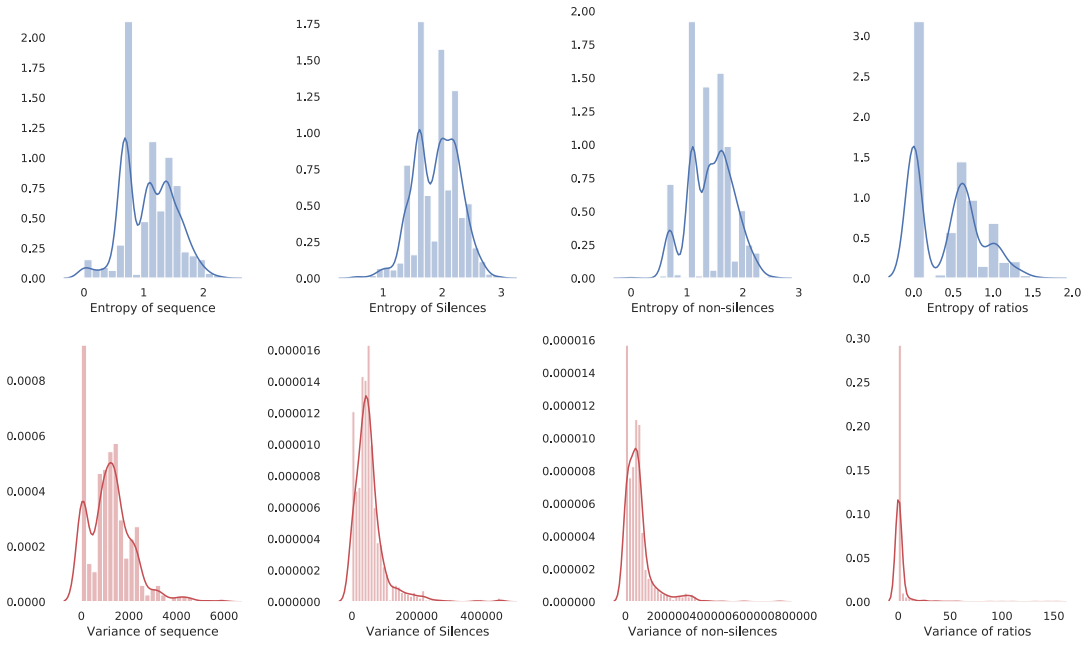

Fig. 3. The distributions of the features used to determine periodicity.

TABLE I
F1 SCORES OBTAINED FOR THE WHOLE SEQUENCE, SILENCE SEGMENTS, NON-SILENCE SEGMENTS AND THE RATIO OS SILENCE TO NON-SILENCE SEGMENTS

	F1 Score	
	Entropy	Variance
Whole sequence	0.6	0.82
Silence segments	**0.85**	0.72
Non-silence segments	0.23	0.62
Ratio of silence to non-silence	0.24	0.26

IV. Results & Discussion

Two experiments were performed on each of the features described in Section II-B. The objective of training is to find the optimal threshold values, ϵ_σ or ϵ_H, for the metric functions $\sigma(x)$ or $H(x)$. As optimization objective, the F1 score is used and its value is maximized.

In Fig. 4 we show how the entropy and variance influence the F1 score for each feature. The best F1 score obtained was from the entropy feature of silence segments. Both the variance and the entropy exhibit similar characteristics except in the case of non-silent segments where the F1 score falls and rises again after an entropy value of 1.1. Overall, the maximum F1 score obtained was from the entropy values of the silence segments.

The results reveal that the entropy of silent segments is better at measuring the periodicity of an audio signal, hence, fax tones, which are highly periodic. In Table I we show the selected threshold values of variance and entropy along with the respective F1 score for each case (whole sequence, silence, non-silence and the ratio of silence to non-silence)

V. Conclusions, Limitations & Future Work

In this study statistical analysis of silence in audio signals is discussed. Specifically call center recordings are considered. The silence segments in audio signals are detected by using the method described in Subsection II-A. The analysis revealed that entropy is a better metric than variance for measuring the periodicity of a signal, yielding an F1 score of 0.85. The variance yielded better F1 scores than the entropy except for the silent segments.

It is worth noting that our audio recordings had varying lengths. This affects the variance and entropy since the number of samples is a factor in determining either of the statistical values. It might be worth exploring the effect of the combination of the features via random forests or similar ensemble machine learning algorithms.

For future studies, we will consider fixed length recordings so as to eliminate the bias that the statistical values have towards shorter or longer call records. We will also test the effect of combining the features. This can be achieved by applying random forests or a mixture of Gaussian models.

References

[1] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, Oct 2013.

[2] A. Zhang, C. Wang, J. Paisley, Q. Wang, and Z. Zhu, "Fully supervised speaker diarization," in *Arxiv*, 2018. [Online]. Available: https://arxiv.org/abs/1810.04719

[3] D. Enqing, L. Guizhong, Z. Yatong, and C. Yu, "Voice activity detection based on short-time energy and noise spectrum adaptation," in *6th International Conference on Signal Processing, 2002.*, vol. 1, Aug 2002, pp. 464–467 vol.1.

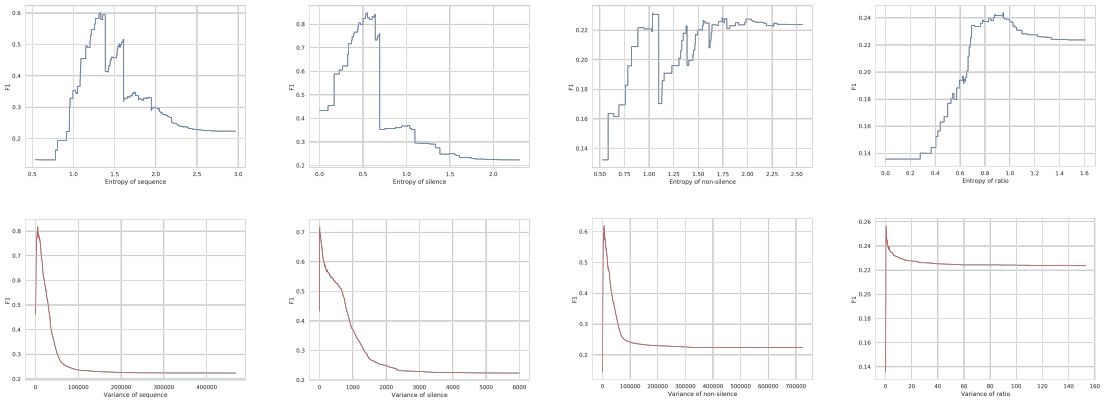

Fig. 4. The variation of the F1 score with computed entropy and variance of silence, non-silence and the ratio of silence to non-silence values.

[4] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *Advances in Information Retrieval*, D. E. Losada and J. M. Fernández-Luna, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 345–359.

[5] N. Chinchor, "Muc-4 evaluation metrics," in *Proceedings of the 4th Conference on Message Understanding*, ser. MUC4 '92. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992, pp. 22–29. [Online]. Available: https://doi.org/10.3115/1072064.1072067

[6] D. Galar and U. Kumar, "Chapter 3 - preprocessing and features," in *eMaintenance*, D. Galar and U. Kumar, Eds. Academic Press, 2017, pp. 129 – 177. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128111536000038