Ses Etkinliği Tespiti için Müşteri-Operatör Görüşmelerinin Makine Öğrenimi Tabanlı İşaretlenmesi
Machine learning-based Annotation of Customer-Operator Conversation Clips for Voice Activity Detection

Leonardo O. Iheme*, Şükrü Ozan*, Erdem Akagündüz†,
* AdresGezgini Inc. Research & Development Center,
Izmir, Turkey
leonardoiheme@adresgezgini.com, sukruozan@adresgezgini.com
†Electrical and Electronics Engineering Department
Çankaya University, Ankara, Turkey
akagunduz@cankaya.edu.tr

***Özetçe*—Bu çalışmada, testlerinde kendi çağrı merkezimizden elde ettiğimiz görüşme kayıtlarını kullandığımız bir ses etkinliği algılama sisteminin geliştirilmesi amaçlanmaktadır. Bu amaçla Naive Bayes sınıflandırıcıyla birleştirilmiş Sesli-Sözcük-Torbası (BoAW) yöntemi uygulanmıştır. Problem, pozitif sınıf olarak konuşma ve negatif sınıf olarak sessizlik ya da arka plan gürültüsü olacak şekilde, ikili bir sınıflandırma problemi olarak formülize edilmiştir. Tüm işlemler, ses kayıtlarından çıkarılan Mel-frekans kepstral katsayı (MFCC) öznitelikleri üzerinde gerçekleştirilmiş; doğruluk skoru ve algılayıcı işletim eğrisi (ROC) olarak sunulan sonuçlar, geliştirilen modelin üst düzey performansını göstermiştir. İlgili sistem veri analizi gerçekleştirmek ve çağrı merkezimizin genel verimliliğini arttırmak üzere çağrı merkezi sunucumuzda konuşlandırılacaktır.**

***Anahtar Kelimeler*—Ses Etkliniği Algılama, Sesli Sözcük Torbası, Mel-frekans kepstral katsayılar, Kümelendirme, Çağrı Merkezi.**

***Abstract*—This study presents the development of a voice activity detection (VAD) system tested on call center telephony data obtained from our local site. The concept of bag of audio words (BoAW) combined with a naive Bayes classifier was applied to achieve the task. It was formulated as a binary classification problem with speech as the positive class and silence/background noise as the negative class. All the processing was performed on the Mel-frequency cepstral coefficients (MFCCs) extracted from the audio recordings. The results which are presented as accuracy score and receiver operating characteristics (ROC) indicate an excellent performance of the developed model. The system is to be deployed within our call center to aid data analysis and improve overall efficiency of the center.**

***Keywords*—Voice Activity Detection, Bag of Audio Words, MFCC, Clustering, Call Center.**

I. INTRODUCTION

Voice activity detection (VAD) has been an active research as well as industry topic for a long time. It has been applied, across a wide range of industries, as a sub-system in several other speech processing systems including real-time speech transmission on the Internet [1] and noise reduction for digital hearing aid devices [2] among others. The problem of detecting speech is not as trivial as it might sound. Challenges arise from the quality of audio being analyzed to the very application specific solutions that exist.

In the literature, the most common methods which have been employed to achieve VAD are: the energy thresholds based approaches [3], pitch detection [4], spectrum analysis [5] or combinations of different features [6]. In recent years, machine and deep learning-based approaches have been applied to VAD tasks. The results obtained, though have been impressive [7], suggest that the models designed for these systems are quite subjective – when tested with data from an unseen corpus, their performances were dramatically degraded.

A major concern for VAD systems is the presence of noise. Most of the VAD algorithms fail when the level of background noise increases [8]. This brings the issue of robustness of the system to fore. For some applications however, this concern may be discounted if the environment is controlled. As is with our application, the issue of noise has not been a hindrance. As we are dealing with telephony data, albeit the low quality of telephone recordings, noise does not hamper the performance of our algorithm. In fact, the presence of noise in the training data is advantageous: it reduces the likelihood of the model over-fitting the data. As noted in [8]: better results are obtained

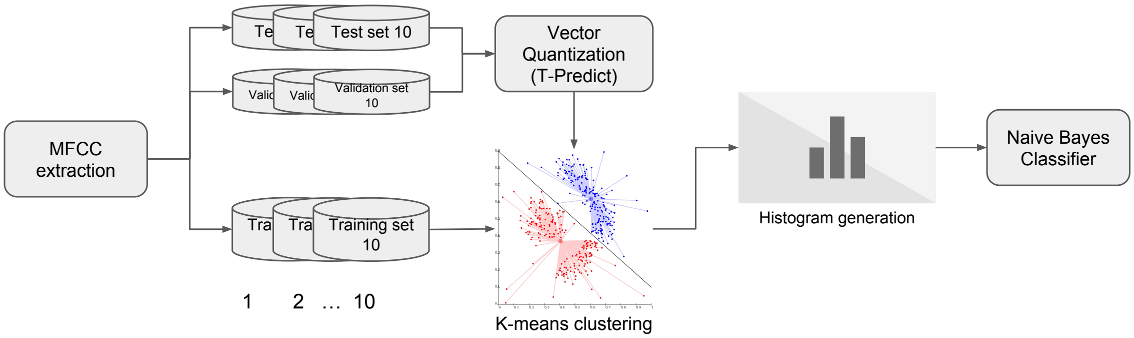


Fig. 1: A Bag of Audio Words (BoAW) Framework for Voice Activity Detection (VAD)

when the machine learning model is trained using a collection of clean and noisy speech records. The Bag of Audio Words (BoAW) approach [9], [10] combined with the Naive Bayes classifier are at the heart of our system. The main advantage of this set-up is that it does not require a lot of training data and the implementation is relatively straightforward.

For call center fraud detection, to decide if a call is fraudulent or not is the ultimate goal. The state-of-the-art systems are based on deep learning algorithms which require a huge amount of training data. Annotation is conventionally carried out by expert human annotators but the process is expensive and the human annotators are liable to be inconsistent. In this study we propose a framework for a machine learning based tool for annotation of voice conversation clips. Our main aim is to construct a semi-automatic tool which can be used for creating a large-scale dataset for call center fraud detection problem.

The next section presents the dataset used in our experiments and in Section III, which is the heart of this paper, the details of the methods are discussed. Our experimental set-up, findings and discussions are presented in Sections IV, V and VI respectively before conclusions future work are outlined in VII.

II. Data

Our corpus is drawn from a central database of telephone calls made over a period of 4 years (2014 - 2018). The database is made up of approximately 1.5 M inbound and outbound call recordings with a total size of about 400 Gb. Calls from our call center are recorded and stored in the compressed Global System for Mobile (GSM) audio format at a sample rate of 8000 Hz. Until the time of writing this article, 60 recordings had been annotated by human experts who had undergone extensive training.

The 60 annotated recordings were randomly selected using SQL queries, with the criteria of having a mean duration and standard deviation of 100 seconds and 30 seconds respectively. Annotation was realized with the Audacity software [11] .

Based on our experts' annotation our data can be summarized in Table I. It shows the duration (in seconds) as well as the distribution of silence and speech (in percentages).

TABLE I: Summary description of data used in experiments

	Duration(s)	Silence(s)	%Silence	speech(s)	%Speech
mean	101.22	17.57	18.97	68.24	65.85
std	31.74	14.06	16.02	30.82	18.04
min	60.60	0.00	0.00	10.96	17.89
max	173.22	58.02	72.49	143.03	90.69

III. Methods

In the usual pipeline for the application of machine learning, after data collection, feature extraction is performed. In the field of digital audio processing, acoustic features are generally extracted. After feature extraction, the next step is training, followed by validation and finally testing. In this section, we will elaborate on these steps. Figure 1 shows the pipeline of our VAD system.

A. *Feature Extraction*

In the realm of digital sound processing, the Mel-Frequency Cepstral Coefficients (MFCCs) have proven to be the most suitable features when it comes to speech recognition. This is because MFCCs accurately model the shape of the vocal tract which is manifested in the envelope of the short-time power spectrum of a signal, ergo, an appropriate model of human sound perception. As this is a basic step, we have left out the theoretical and implementation details. Our system makes use of 12 MFCCs as well as the log energy, extracted using the Librosa library [12].

B. *Bag-of-Audio-Words*

The concept was motivated by the Bag of Words algorithm from document classification where the frequency of occurrence of each word is used as a feature for training a classifier [13]. The issue of variable-length audio files can be resolved by grouping similar features together in a fixed length histogram. We predefine the number of ”words” for the codebook using the k-means clustering algorithm. The centroids of the resulting clusters are taken as the codewords. Each MFCC vector is then replaced by the index of the codeword nearest to it. Finally, for every file, a histogram of codewords is generated.

The number of codewords (codebook), i.e. *k*, needed for optimum performance is a trade-off between how descriminative or general we want our ”vocabulary” to be. A large codebook is likely to assign similar sounds to different clusters thus being more descriminative. However, with a smaller codebook, we stand the risk of assigning dissimilar sounds to the same cluster. In this work, we examined codebook sizes between two and 1500.

C. *Naive Bayes Classifier*

In our work, we exploit the success of the naive Bayes classifier as applied to text classification tasks. It is advantageous for its simplicity and proven effectiveness in handling real world problems. Specifically, we apply a Gaussian naive Bayes which assumes that the values associated with each class in the training data are distributed according to a Gaussian distribution.

Before feeding the histograms to the classifier, it is common practice to perform Laplace smoothing to account for unseen codewords, ergo 0 posterior probabilities. Concretely, we increment each value in the histogram by one and divide by the sum of the codewords and the total number of words.

IV. Experimental Set-up

We split our data into a 80:10:10 ratio: training, validation and testing respectively. We performed a 10-fold cross-validation to determine the optimum parameter values that maximize the prediction accuracy of our model. After validating the parameters, we predicted the labels of the samples in our test set (unseen data).

Since our goal was to detect speech/silence, we needed to classify a group of MFCCs whose combined duration corresponded to meaningful speech/silence. We call this parameter *T-predict*. Briefly, we collect a number of adjacent MFCCs that have the same label in one chunk. Chunks that contained a mixture of labels or chunks where the number of MFCCs was less than *T-predict* were discarded.

The parameters which were optimized are the codebook size, *k* and *T-predict*. We performed a grid search for *k* between 2 and 1500 and *T-predict* between 2 and 70.

A. *Evaluation*

Primarily, our results were evaluated by computing the prediction accuracy of our model. Furthermore we examined the performance of the model by plotting the ROC and calculating the AUC.

V. Results

We trained our model by optimizing the hypeparameters via cross validation with the mean prediction accuracy as the criterion. The effect of *T-predict* and *k* on the prediction accuracy is depicted in Figure 2. We obtained a mean prediction accuracy of 0.99 and a standard deviation of 0.03 with *T-predict* = *50* and *k* = *2* at best and 0.82±0.05 with *T-predict* = *2* and *k* = *1500* at worst.

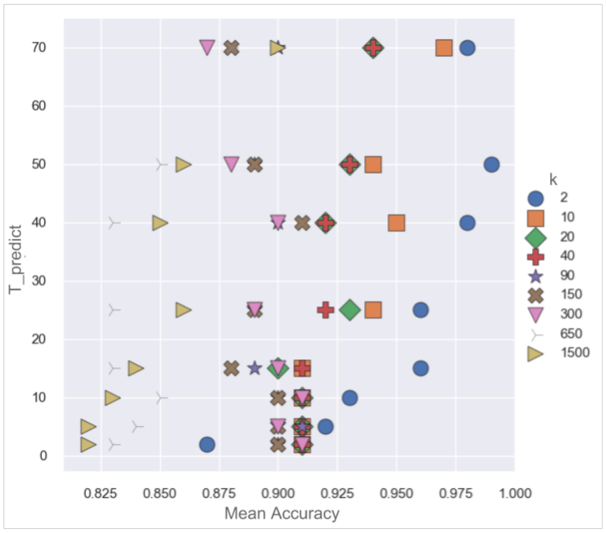


Fig. 2: Result of Cross Validation grid search over *T-predict* and the number of clusters *k*

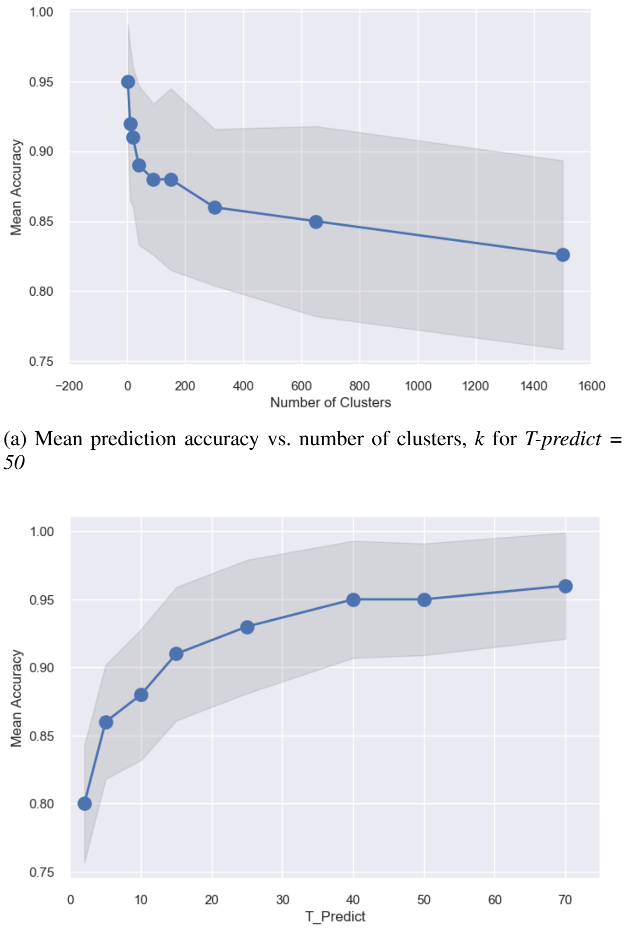
On the test set, we also examined the relationship between the prediction accuracy and the two hyperparamaters by keeping one constant at its optimum and varying the other. In Figures 3 we show how the prediction accuracy is related to *k* and *T-predict*. The figures show that our model achieved a prediction accuracy of 0.96 (standard deviation of 0.04). To further analyze the performance of the model, we explored the measure to which it can distinguish between speech and silence. The plots of the ROC curves and AUC computation as depicted in Figure 4 provide this insight.

VI. Discussion

A salient pattern can be observed from the results we have presented. The prediction accuracy seems to be directly proportional to *T-predict* and inversely proportional to the size of the codebook, i.e. *k* (Figure 3). Since we are dealing with a binary classification problem, it intuitively makes sense to arrive at an optimum codebook size of two. It implies that our codebook is quite descriminative of the two classes (speech and silence/noise). The optimum value of *T-predict* obtained reveals the duration of time required to classify a sound segment as speech or silence/background. especially since we are not dealing with classification of formants, a relatively large *T-predict* was expected.

By keeping *k* constant and varying *T-predict*, and vice-versa the same trend can be observed on the test set. Figures 3a and 3b give us a sense of how the mean accuracy is related to each of these parameters. The Figures also show us that the standard deviation does not exceed 0.05 (at worst).

As the accuracy score may be mis-leading on an imbalanced data set, we computed the ROC score and plot in Figure 4. The excellent results obtained can be attributed to ”bagging” in BoAW and the Laplace smothing before applying naive Bayes classifier. Figure 4 shows that the system can almost perfectly predict the correct class of a previously unseen sample.



(a) Mean prediction accuracy vs. number of clusters, *k* for *T-predict* = *50*

(b) Mean prediction accuracy vs. *T-predict* for *k* = *2*

Fig. 3: Plot of prediction accuracy vs. the best hyperparameter values. Gray areas denote the standard deviation among experiments.

VII. Conclusion and Future Work

In this work, we have described as well as presented the results of testing our bag of audio words-naive Bayes VAD system. The algorithm performed excellently on our call center data, reaching a prediction accuracy of up to 96%.

We gathered that, because we wanted to detect segments of speech and silence/background noise, the prediction accuracy is directly proportional to the number of MFCC vectors (*T-predict*) being classified. The prediction accuracy peaked at *T-predict* $= 50$. On the other hand, prediction accuracy had an inversely proportional relationship with the codebook size, peaking at $k = 2$. This revelation satisfied our intuition since the classification task was a binary one.

To improve on the current semi-automatic audio data labelling tool, we will focus on the detection of other attributes of a telephone call including music, tones and different speakers. By properly formulating a multi-class classification problem, we believe that it can be achieved.

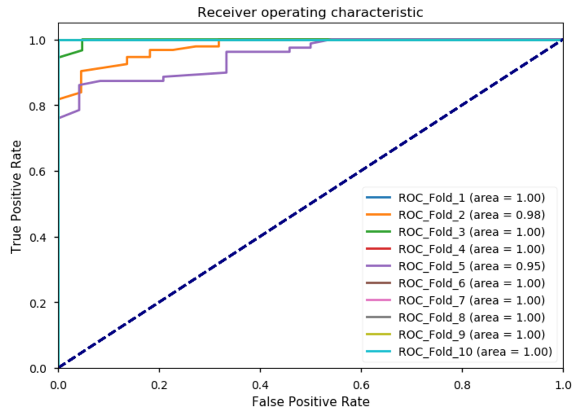


Fig. 4: Receiver Operating Characteristics (ROC) and Area Under the Curve (AUC) for 10 folds of the test set

References

[1] A. Sangwan, M. C. Chiranth, H. S. Jamadagni, R. Sah, R. V. Prasad, and V. Gaurav, "Vad techniques for real-time speech transmission on the internet," in *5th IEEE International Conference on High Speed Networks and Multimedia Communication (Cat. No.02EX612)*, July 2002, pp. 46–50.

[2] K. Itoh and M. Mizushima, "Environmental noise reduction based on speech/non-speech identification for hearing aids," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, April 1997, pp. 419–422 vol.1.

[3] K.-H. Woo, T.-Y. Yang, K.-J. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, Jan 2000.

[4] R. Chengalvarayan, "Robust energy normalization using speech/nonspeech discriminator for german connected digit recognition." in *EUROSPEECH*, 1999.

[5] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 109–118, Feb 2002.

[6] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 478–482, July 2000.

[7] S. Tong, H. Gu, and K. Yu, "A comparative study of robustness of deep learning approaches for vad," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5695–5699.

[8] J. Ramirez, J. M. Gorriz, and J. C. Segura, "Voice activity detection. fundamentals and speech recognition system robustness," in *Robust Speech*, M. Grimm and K. Kroschel, Eds. Rijeka: IntechOpen, 2007, ch. 1. [Online]. Available: https://doi.org/10.5772/4740

[9] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification." in *INTERSPEECH*. ISCA, 2012, pp. 2105–2108. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2012.htmlPancoastA12

[10] A. Plinge, R. Grzeszick, and G. A. Fink, "A bag-of-features approach to acoustic event detection," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 3704–3708.

[11] "Audacity® software is copyright© 1999-2018 audacity team. it is free software distributed under the terms of the gnu general public license. the name audacity® is a registered trademark of dominic mazzoni." [Online]. Available: https://audacityteam.org/.

[12] B. McFee et al., "Librosa library v.0.6.2," Aug. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1342708

[13] M. McTear, Z. Callejas, and D. Griol, *The Conversational Interface: Talking to Smart Devices*, 1st ed. Springer Publishing Company, Incorporated, 2016.