Confidence-Based Scoring: A Useful Diagnostic Tool for Detection Tasks

TJ Tsai[1,2], *Adam Janin*[2]

[1]EECS Department, University of California at Berkeley, Berkeley, CA, USA
[2]International Computer Science Institute, Berkeley, CA USA

tjtsai@eecs.berkeley.edu, janin@icsi.berkeley.edu

Abstract

This paper uses an unconventional analysis as a tool to diagnose the problems with three different speech activity detection systems. The unconventional analysis is to score the frames in an audio file in order of confidence, starting with the frame that we have the most confidence in and progressing towards less and less confident frames. By keeping track of the cumulative number of errors, we can determine how the errors are distributed across the data. Using speech activity detection on highly degraded audio as a case example, we show how this simple analysis can yield useful insight into system performance. In our case example, we use the analysis to establish that (1) a small percentage of the frames account for a lion's share of the errors, (2) three different systems perform very poorly on the same small subset of 'hard' data, and (3) the 'hard' data is primarily characterized by its proximity to speech-nonspeech boundaries. Through follow-up analyses, we show that the boundaries are 'smoothly' hard, and that scoring collars alone are not enough to handle the problem. Through this case example, we demonstrate the utility of confidence-based scoring as a general diagnostic tool for detection tasks on time-series data.

Index Terms: confidence-based scoring, speech activity detection

1. Introduction

Traditionally, detection systems are characterized by a receiver operation characteristic (ROC) or detection error tradeoff (DET) curve. The ROC/DET curve shows the tradeoff between false alarm errors and miss detect errors. This paper introduces a simple analysis method that allows us to gain deeper insight into system performance by showing how errors are distributed across the data. For audio detection tasks, the analysis would consider frames in each audio file in order of confidence, starting with the frame with highest confidence and progressing towards frames with less and less confidence. By keeping track of the cumulative number of errors, we can determine how the errors are distributed across the data when ordered by confidence. Using speech activity detection on highly degraded audio as a case example, we apply this simple analysis to three existing systems in order to diagnose the problems with these systems. Through this case investigation, we demonstrate the utility of confidence-based scoring as a general diagnostic tool for detection tasks on time-series data.

We now turn our attention to the specific case example at hand: speech activity detection (SAD). By applying this analysis to three different SAD systems on RATS data, we establish three main points. First, a small fraction of the data accounts for a large fraction of the errors. Second, the three SAD systems all perform very poorly on this small fraction of 'hard' data. Third, this small fraction of 'hard' data is primarily characterized by its proximity to speech-nonspeech boundaries. We show that the results we observe are not simply a consequence of ground truth inaccuracy, but rather a steady, observable progression of data becoming increasingly difficult as we move closer to the boundary. After explaining the experimental setup in section 2, we demonstrate the first two points in section 3 and establish the third point in section 4. Section 5 summarizes and concludes the work.

2. Experimental Setup

We explain the experimental setup in three parts: the system descriptions, the data, and the analysis methodology.

2.1. System Descriptions

In this section we describe the three SAD systems. Since our goal in this paper is not to advocate a particular approach to the task of SAD but rather to demonstrate the utility of confidence-based scoring as a diagnostic tool, the fine details about these three systems are not of primary importance. Therefore, the descriptions given here are limited to a high level overview rather than an in-depth explanation, so that more space can be given to the analysis sections. The most important piece of information in this section is to simply note that the three systems are very different from each other – they have very different decoding algorithms and very different features. In the remainder of this section, we briefly describe these three systems and include appropriate references and background work for the interested reader.

The first system is a two-state hidden markov model with MFCCs. The two states correspond to speech and nonspeech, and each state is modeled as a mixture of gaussians. The training and decoding algorithm for this system will be abbreviated as HMM-GMM for the remainder of this paper. The MFCC features are 39 dimensional and include first and second order derivatives. This first system is typical of an approach that might be used as a reference baseline.

The second system is an HMM-GMM with spectrotemporal modulation features. Spectrotemporal modulation features have been explored in the last decade as a more flexible and powerful feature representation for various speech-related tasks. The main idea of these features is to consider the spectrogram as a two-dimensional image and to measure the response of various two-dimensional filters at different patches in the spectrogram. These features are biologically motivated (see [1], for example), and seek to approximate the types of auditory stimuli that cause neurons in the brain to fire. In our experiments, we used the set of two-dimensional Gabor wavelet filters described in [2] and [3], which results in a 449 dimensional feature vector. For recent work in applying spectrotemporal modulation features to the SAD task, see [4], [5], and [6].

The third system is an approach that dynamically combines a set of weak classifiers using voicing-related features. It treats each feature as a weak classifier and assigns weights to the weak classifiers in a dynamic fashion. Unlike approaches such as Adaboost where the weights of the classifiers are determined during training and fixed during prediction, this system determines the weights dynamically by considering how confident each weak classifier is in its current prediction. It adopts the approach in [7] referred to as dynamic selection, in which the most confident weak classifier is given all the weight and thus completely determines the prediction. The features for this system consist of a base voicing feature and a set of 220 derived voicing features. The base feature is a probability of voicing estimate by a subband autocorrelation pitch tracker, which is described in [8]. Using this probability of voicing at every frame as a base feature, we then derived a family of features by calculating statistics on windows of various sizes. The statistics we considered were the minimum, the maximum, and various quantiles in between (where, for example, the 50% quantile would correspond to the median). We considered windows up to 2 seconds long. A complete description of this system is given in [9]. This system was designed based on the metaphor of an economics marketplace, and will be referred to as the economics approach.

2.2. Data

Our experiments used data from the DARPA RATS program. The data consists of conversations recorded over various radio transmission links. In general, the audio data is very noisy and contains highly non-stationary noise, including high energy non-transmission regions. Due to ground truth label integrity issues, we randomly selected 1 minute segments and manually verified the labels, throwing out any segments that had poor labels. Our final data set consisted of 523 training segments and 324 evaluation segments. For more information on the data and on other SAD approaches proposed for this data set, see [10], [11], and [12].

2.3. Analysis Methodology

Our analysis is based on what we will call an error trajectory. SAD systems are traditionally characterized by ROC curves, which show the tradeoff between false alarm (FA) and miss detect (MD) errors. An error trajectory examines a single point on the ROC curve and shows how the FA and MD errors are distributed across the data. (Note that, since we would like to show the breakdown of total errors into its constituent parts, we prefer the linear axes of the ROC plane over the non-linear axes of the DET plane.) To compute an error trajectory, we consider the frames in each audio file in order of confidence, starting the with the frame that we have the most confidence in and progressing towards less and less confident frames. By keeping track of the cumulative number of errors, we can determine how errors are distributed across the frames. An error trajectory thus begins at the (0,0) point in the ROC plane (before it has scored any frames), and it ends at the specified point on the ROC curve (after it has scored all frames). The error trajectory will serve as the foundation for our analysis.

In addition to explaining the conceptual idea of an error trajectory, there are a few practical considerations that are important to mention. One very useful visualization technique is to demarcate points along the error trajectory at regular intervals. In the error trajectories we show in this paper, plotted points show milestones in 5% increments. In other words, each successive segment between two plotted points shows the amount of error contributed by another 5% of the frames in an audio file (averaged across the evaluation data set). These demarcated points make it easy for us to see, for example, how much error was contributed by the most confident 20% of frames. Another important consideration is to point out that we exclude frames within scoring collars. When comparing system hypotheses to ground truth labels in an SAD task, it is customary to exclude frames that are very close to speech-nonspeech boundaries in order to accommodate for ground truth label inaccuracies. In the error trajectories we present, we use a 200 ms speech scoring collar and a 500 ms nonspeech scoring collar, which are the collar sizes specified in the official RATS SAD evaluation. This means that the 200 ms on the speech side of each boundary and the 500 ms on the nonspeech side of each boundary will not be scored. Lastly, it is important to point out that the error trajectories for two different SAD systems are not directly comparable, since the two systems will yield different confidence orderings of the frames. For this reason, we adopt the technique of taking the ordering of frames from one system and applying it to all other systems. By doing this, we make the trajectories directly comparable, since corresponding segments will thus refer to the exact same subset of frames. Trajectories whose ordering is adopted from a different system will be referred to as imputed error trajectories.

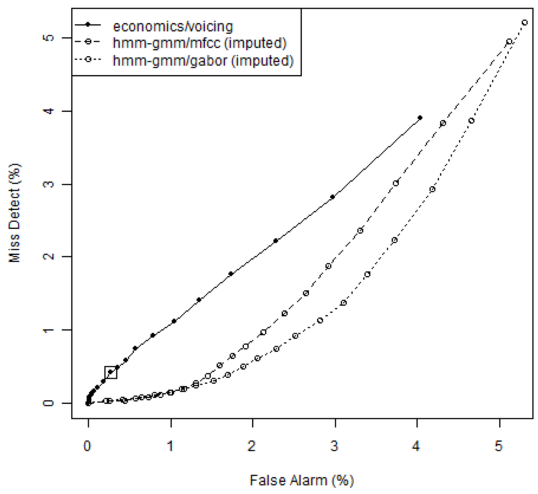

Figure 1: Error trajectories for three different speech activity detection systems showing how system errors are distributed across the data. Note that the confidence ordering for all three trajectories has been adopted from the economics system, so corresponding segments of different trajectories refer to the exact same subset of data.

3. Results

Figure 1 shows the error trajectory for the economics marketplace system and the imputed error trajectories for the two HMM-GMM systems. This means that the ordering of frames for all three trajectories is adopted from the economics system. There are two important things to notice about this figure.

First, a small subset of frames accounts for a lion's share of the errors. The spacing of the demarcated points in the solid line is very dense at the beginning of the trajectory and comparatively very sparse towards the end of the trajectory, indicating that we are performing very well when we're confident and very poorly when we're not confident. In fact, the 50% of the data that the economics system is most confident in accounts for only 7% of the total FA errors and 11% of the total MD errors (this is denoted by the boxed point on the solid line). In contrast, the 13% of the data that the economics system is least confident in accounts for 50% of the total number of both FA and MD errors (this corresponds roughly to the point at 2% FA and 2% MD). Recall that these results have already excluded very generous scoring collars, so the extremely poor performance on the least confident frames cannot simply be attributed to ground truth inaccuracy. It appears that there is a small set of legitimately scored frames that is responsible for most of the errors.

Second, all three systems agree on what is hard. Note that the last few segments in all three trajectories are very long, showing that all three systems perform very poorly on the same subset of data (recall that the three trajectories are all using the ordering adopted from the economics system, so corresponding segments in different trajectories are directly comparable). Given that the three systems are fairly diverse in their decoding algorithms and features, a reasonable explanation for this phenomenon is that this small subset of frames is inherently difficult to classify – regardless of the type of feature or decoding algorithm. While we can only directly show that these three particular systems agree on what is hard, we can hypothesize that perhaps most if not all SAD systems would agree that this subset of frames is the most difficult to classify.

The analysis above suggests that the most reasonable way of improving our SAD systems is to focus on the 13% of least confident frames, which contributes 50% of our total errors. Who are these 13%? Answering this question is the focus of the next section.

4. Discussion

In this section we will show, through a series of follow-up analyses, that the frames that contribute the most errors are those near speech-nonspeech boundaries. We also demonstrate that this is not merely an artifact of ground truth inaccuracy, but rather a steady, observable progression of the data becoming increasingly hard to classify (in a statistical sense) as we move closer to the boundary. We rely on three different methods of follow-up analysis to demonstrate these claims.

The first analysis is to consider the distance from every frame to the nearest speech-nonspeech boundary (in ground truth). Once we have computed this distance for every frame, we can then consider the distribution of these distances for various subsets of frames. Figure 2 shows the distributions for 4 different subsets: the 25% of highest confidence frames, the 25% of lowest confidence frames, and the two quartiles in between. As the confidence level decreases, there is a clear progression of the data being concentrated more and more closely to the boundaries. Note that this progression is not just a characteristic of the small subset of hard frames, but a steady progression that we observe from the very beginning with the highest confidence frames.

The second analysis is to investigate the temporal sequence of frames when ordered by confidence. Consider an audio file with 6000 frames and imagine the frames as a sequence of boxes next to each other. Let's say that the most confident frame in this file is frame 3500 and it is a prediction for speech, so we fill in box 3500 with "sp". The next most confident frame is frame 2000 and it is a prediction for nonspeech, so we fill in box 2000 with "ns". We continue this process until all frames have been filled in. Every time we fill in a box, we answer three questions about that box: Does this frame extend an already existing "island" of frames of the same type? Does this frame start a new island? Is this frame adjacent to an island of the opposite type? Frames that satisfy these rules are considered to be extenders, starters, and opposers, respectively. In the example above, the first two frames would be starter frames, since they both start an island (i.e. in each case the two adjacent frames have not been scored yet). If the third most confident frame was frame 3501 and it was a prediction for speech, it would be an extender since it extends the existing speech island at frame 3500. (If it was a prediction for nonspeech, it would be an opposer instead, since it effectively ends the rightward growth of the adjacent speech island.) Note that a frame can simultaneously be an extender and an opposer if it fills in a box between a speech island and a nonspeech island. The notions of starter, extender, and opposer frames is depicted in figure 3.

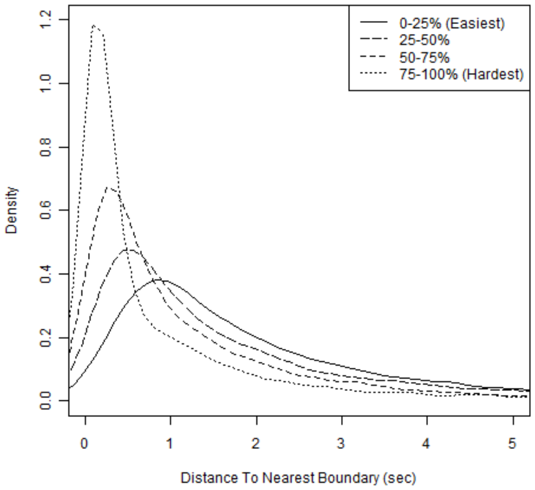

Figure 2: Distribution of distance to nearest speech-nonspeech boundary grouped by relative confidence level. As the confidence level decreases, frames tend to be concentrated more and more closely to the boundaries.

By keeping track of the roles that each frame plays, we can gain an intuitive understanding of the temporal sequence of frames when ordered by confidence. Table 1 shows the fraction of frames that fulfill these different roles when the frames are divided into 4 subsets of confidence quartiles. We see that approximately 90% of the frames are extending islands of the same type, and the other 10% of frames are starting new islands. There are almost no opposers until we get to the least confident frames, and even then they occur very sparingly. Let's describe the mental picture that these numbers paint. The highest confidence frames start in the middle of speech or nonspeech segments. As we progress to less and less confident frames, these islands of scored frames extend and grow larger. Occasionally, a new island will appear, and it will extend and grow as well. Islands of the opposite type (speech vs nonspeech) do not touch each other until we reach the lowest confidence levels. It is in

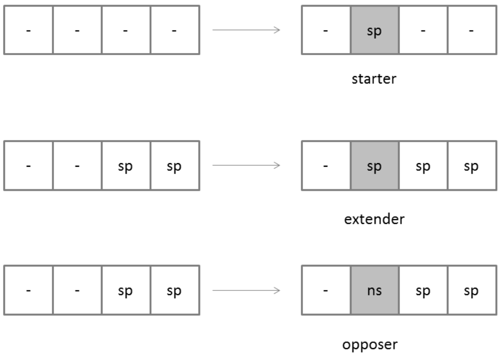

Figure 3: Examples of starter, extender, and opposer frames. When considering frames in an audio file in decreasing order of confidence, knowing what fraction of frames satisfy these various roles helps to gain insight into the data.

these regions – the locations of contact between speech islands and nonspeech islands – that we accumulate the most errors.

Difficulty Quantile	Starters	Extenders	Opposers
0-25%	7.0%	93.0%	0.0%
25-50%	8.7%	91.3%	0.0%
50-75%	9.5%	90.4%	0.1%
75-100%	11.7%	87.2%	2.6%

Table 1: Percentage of starter, extender, and opposer frames. The first row refers to the 25% of frames with highest confidence, while the last row refers to the 25% of frames with the lowest confidence.

The third analysis is to consider segments of low confidence frames, to stitch the corresponding audio segments together, and to listen to them. Though this analysis is not quantitative, it is nonetheless useful to share our qualitative observations. The hard speech data consists of lots of beginning and ends of words, transient sounds, and breaths or exhalations. The hard nonspeech data also contained a lot of transient sounds, such as transmission onsets/offsets or suddenly changing noise conditions. These qualitative observations are consistent with the explanations suggested by the previous two analysis methods.

These three confidence-based analysis methods help to answer the question: who are the 13%? The frames that contribute the lion's share of errors are those near speech-nonspeech boundaries. The analyses show that there is a steady, observable progression of data becoming harder as we move closer to the boundaries. One important practical implication of this conclusion is that researchers who report results on SAD tasks should always specify the size of their scoring collars. Since the majority of errors occur near speech-nonspeech boundaries, the size of the scoring collars will greatly affect the results. For example, when we reduce the scoring collars in our experiments down to 50 ms, we observed that error rates increased by 30-70%. In our reading of SAD literature, the scoring collars are rarely (if ever) mentioned, and this omission makes results difficult to interpret. Another potential implication of these analyses is to reconsider the general approach to SAD on highly degraded audio. Since we know that most or all systems will perform very poorly on the region around the boundaries, approaching the problem as a boundary estimation problem rather than a frame-level classification may be a more fruitful avenue of exploration.

5. Conclusions

This paper uses an unconventional confidence-based scoring analysis as a useful diagnostic tool to gain deeper insight into system performance and data. Using speech activity detection in highly degraded audio as a case investigation, we apply this analysis method to demonstrate three things. First, a small fraction of the data is responsible for a large fraction of the errors. Second, three different speech activity detection systems all agree on what data is hard. Third, the hard data is primarily characterized by its proximity to speech-nonspeech boundaries. Through several follow-up analyses, we show that this is not merely an artifact of ground truth inaccuracy, but rather a steady, observable progression of declining system performance as we move closer to the boundaries. Through this case example, we show how this analysis can yield useful insights into the data. The analyses in this paper would be applicable to any detection task on time-series data.

6. Acknowledgements

Thanks to Nelson Morgan for his feedback and comments on this paper. This material is based on work supported by DARPA under contract no. D10PC20024. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of DARPA or its contracting agent, the US Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch.

7. References

[1] Mesgarani, N. and Shamma, S., “Speech Processing with a Cortical Representation of Audio”, Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5872 - 5875, 2011.

[2] Meyer, B. T., Ravuri, S. V., Schadler, M. R., and Morgan, N., “Comparing Different Flavors of Spectro-Temporal Features for ASR”, in Proc. of Interspeech 2011.

[3] Tsai, T. and Morgan, N., “Longer Features: They do a speech detector good”, in Proc. of Interspeech 2012.

[4] Mesgarani, N., Slaney, M. and Shamma, S. A., “Discrimination of Speech From Nonspeech Based on Multiscale Spectro-Temporal Modulations”, in IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 3, May 2006.

[5] Markaki, M. and Stylianou, Y., “Discrimination of speech from nonspeech in broadcast news based on modulation frequency features”, Speech Communication 53, pp. 726 - 735, 2011.

[6] Bach, J. H., Anemuller, J. and Kollmeier, B., “Robust speech detection in real acoustic backgrounds with perceptually motivated features”, Speech Communication 53, pp. 690-706, 2011.

[7] Tsymbal, A. and Puuronen, S., “Bagging and boosting with dynamic integration of classifiers”, Principles of Data Mining and Knowledge Discovery, pp. 195-206, 2000.

[8] Lee, B. S. and Ellis, D., “Noise Robust Pitch Tracking by Subband Autocorrelation Classification”, Proc. of Interspeech 2012.

[9] Tsai, T. and Morgan, N., “Speech Activity Detection: An Economics Approach”, Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2013.

[10] Walker, K. and Strassel, S., “The RATS Radio Traffic Collection System”, Odyssey 2012-The Speaker and Language Recognition Workshop 2012.

[11] Ng, T., Zhang, B., Nguyen, L., Matsoukas, S., Vesely, K., Matejka, P., Zhu, X., and Mesgarani, N., "Developing a speech activity detection system for the darpa rats program", Proc. of Interspeech 2012.

[12] Thomas, S., Mallidi, S. H., Janu, T., Hermansky, H., Mesgarani, N., Zhou, X., Shamma, S., Ng, T., Zhang, B., and Nguyen, L., "Acoustic and Data-driven Features for Robust Speech Activity Detection", Proc. of Interspeech 2012.