INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Minimizing Annotation Effort for Adaptation of Speech-Activity Detection Systems

Luciana Ferrer[1], *Martin Graciarena*[2]

[1]Departamento de Computación, FCEyN, Universidad de Buenos Aires and CONICET, Argentina
[2]Speech Technology and Research Laboratory, SRI International, California, USA

lferrer@dc.uba.ar

Abstract

Annotating audio data for the presence and location of speech is a time-consuming and therefore costly task. This is mostly because annotation precision greatly affects the performance of the speech-activity detection (SAD) systems trained with this data, which means that the annotation process must be careful and detailed. Although significant amounts of data are already annotated for speech presence and are available to train SAD systems, these systems are known to perform poorly on channels that are not well-represented by the training data. However obtaining representative audio samples from a new channel is relative easy and this data can be used for training a new SAD system or adapting one trained with larger amounts of mismatched data. This paper focuses on the problem of selecting the best-possible subset of available audio data given a budgeted time for annotation. We propose simple approaches for selection that lead to significant gains over naïve methods that merely select N full files at random. An approach that uses the frame-level scores from a baseline system to select regions such that the score distribution is uniformly sampled gives the best trade-off across a variety of channel groups.

Index Terms: speech-activity detection, adaptation, annotation, active learning

1. Introduction

Speech-activity detection (SAD) is an essential step in most speech-processing tasks, such as speech, speaker and language recognition. In these cases, algorithms usually rely on a SAD stage to filter out non-speech, keeping only the regions containing speech, the information relevant for the task. SAD is also a task in its own right in cases in which large amounts of audio are collected and searched for speech, which is later given to a human for manual analysis.

Currently, the best-performing SAD systems are based on different kinds of deep neural networks (DNNs) trained on large amounts of data [1, 2, 3, 4]. For optimal performance, the training data should be somewhat matched in terms of acoustic characteristics to the data that the model will be applied to. When mismatch is present, performance can degrade to unusable levels, depending on the nature and degree of the mismatch. In these cases, as we will show, using even a small amount of matched labeled data to adapt the original model to the new channel characteristics can greatly improve performance.

Although collecting audio data from a new channel is usually easy and cheap, annotating this data for the presence of speech for training or adapting a SAD system is not an easy task. Our experiments indicate that an error on the order of a tenth of a second in the determination of the start and end times of the speech regions may result in significant performance degradation of the trained model. Hence, the boundaries between speech and non-speech should be precisely labeled and even short pauses in between spurts of speech should be labelled as non-speech. For this reason, the process of annotating data for training a SAD system is laborious and time-consuming.

In this paper, we assume a scenario where large amounts of labeled data are available for training an initial SAD system from a set of channels different from those used for testing. We also assume that a certain amount of unlabelled data is available for each test channel. The goal is then to select a subset of this data for manual annotation, which can then be used for adapting the system to the channel or acoustic conditions of interest. Given a certain budgeted time for selection, the challenge is selecting the data as intelligently as possible to optimize the performance of the final adapted system. This work was motivated by the SAD evaluation organized as part of the Robust Automatic Transcription of Speech (RATS) program of the Defense Advanced Research Projects Agency (DARPA) .

In a general sense, our goal coincides with that of the field of active learning (see, for example, [5] for a review of active learning). In active learning, a certain amount of unlabeled data is available for training. Iteratively, models are trained with increasing subsets of this data, which are selected for their potential to improve the training system performance at each stage. Several different criteria are used for selecting data at each stage, including selection of the samples with least certain decisions, samples with less agreement across a variety of models, samples that would change the model the most if used for training, and so on.

The SAD task, though, deviates from the usual active learning setup in that the unit or samples available for labelling are not predefined. While we could consider the samples to be the waveforms on one extreme, or the frames (10 millisecond snippets of audio) at the other extreme, neither of these options is ideal. Frames are too short to be labeled in isolation by a human for the presence of speech, while full waveforms might be redundant when considered fully, depending on their length. Ideally, we would like to give annotators small regions of audio of a certain minimum duration (making annotation efficient), optimally selected from across all available waveforms (making them most useful for adaptation). These regions can potentially be located anywhere within the available signals, may be of variable length and contain both speech and non-speech. Further, in most SAD systems, regions are not assigned a single score measuring the system's certainty about the decisions made within that region. Rather, one score per frame is usually calculated. Hence, standard active learning techniques that rely on a score computed by a previously-trained model to determine the usefulness of a sample (a region in our case) cannot be directly applied to SAD. These issues make the application

http://dx.doi.org/10.21437/Interspeech.2016-247

of standard active learning techniques to SAD a non-trivial task.

In this work, we take first steps into the application of active learning ideas for the SAD task. We focus on simple selection approaches that, for simplicity, assume that a single stage of annotation is performed, rather than an iterative process. At this stage, the model trained with mismatched data can be used for selecting the optimal regions for annotation. The selected regions and their annotations are then used for adaptation of this initial model. We show results for varying durations budgeted for annotation, analyzing the effect of the selection algorithm and the adaptation parameters.

2. DNN-Based SAD

DNN-based SAD systems are currently the state of the art [1, 2, 3, 4]. These systems use as model a DNN trained to predict the posterior of the speech and non-speech classes at the output layer. The posteriors are converted into log-likelihood ratios (LLRs) by using Bayes rule, assuming equal priors for both classes. In a final step, these LLRs are smoothed by averaging their values over a rolling window, typically 31 to 71 frames long. The final SAD decisions are made by thresholding these LLRs, with a threshold chosen based on the desired operating point. For some applications, the resulting speech regions (contiguous frames with LLR value above the threshold) are padded with a certain number of frames on each side. This padding reduces the amount of missed speech near the detected speech regions while potentially increasing the false alarm rate.

Currently the best-performing SAD systems use a DNN architecture including some short long-term memory (SLTM) layer or layers [6]. In this work, though, we use the standard feed-forward DNN architecture, with only two hidden layers of sizes 500 and 100. Training is performed by using several iterations of the backpropagation algorithm, with small mini-batches and cross-entropy as the error metric. Iterations are stopped when the performance in a held-out cross-validation set stops improving. The held-out set is determined as a random 15% of the available training waveforms. In our work we use mel-frequency cepstral coefficients (MFCCs) concatenated over 31 frames to include contextual information as input to the DNN.

In this work, we assume a scenario where a large amount of labelled data is available for training the SAD system, but this data is mismatched to the channels of interest. For these channels, we only have small amounts of labelled data, dynamically selected from within a larger set of unlabelled data by using the different algorithms described in Section 3. In this scenario, we have found that an efficient way of using this small amount of data is to adapt the big mismatched model to it. The adaptation is performed by doing additional iterations of back propagation, using the mismatched model to initialize the parameters at the first iteration. Convergence, in this case, is determined on all the available data, because leaving even a small percent out for cross-validation degrades final model performance.

The key for getting good performance with this method when using small amounts of adaptation data is to control the amount of change allowed in the adaptation iterations by using regularization. The regularization is done by adding a term to the objective function given by the L2 norm of the model parameters with respect to the parameters in the previous iteration. This term is weighted by a tunable factor that we call the regularization factor. In our experiments, we explore a strategy for fixing this value as a function of the amount of annotated data, which results in significant gains with respect to using a fixed value optimized for a certain duration.

3. Selecting Regions for Annotation

Annotating audio signals for the presence and location of speech for use as SAD system training data is a time-consuming task. Annotating approximate speech regions is generally relatively easy, but selecting the exact bounds requires more careful signal analysis. Unfortunately, the annotation precision directly influences the quality of current DNN-based systems.

To assess the effect of annotation error on system performance, we corrupted the annotations provided by the RATS program by padding each speech region with 10 or 20 frames (or 0.1 and 0.2 seconds) to each side. This padding emulates what is likely to happen in quick annotation of speech regions: short snippets of non-speech (breathing, hesitations) are simply absorbed into the surrounding speech regions and the start and end times are stretched slightly into the non-speech regions around them. Training the model with annotations corrupted by 10 frames of padding results in relative degradation between 10 and 18% with respect to the model trained with the original annotations, over the four channel groups described in Section 5. When corrupting the annotations with 20 frames of padding, the degradations increase to 18 to 35%. These results clearly indicate the potentially large effect of the annotation error on SAD performance. This finding justifies the large effort required to carefully annotate data for SAD system training (or adaptation).

As mentioned above, in this work we assume a scenario where an initial model is trained on data from channels different from the one present in testing. We also assume that a relatively large set of unlabelled samples from the channel of interest is available for adaptation. The goal is then to select a subset of this data of a certain budgeted duration B (measured in minutes) for annotation and then adaptation, aiming to optimize the selection to obtain the best possible adapted model. In the following sections we explain the different approaches we use for selecting regions for annotation.

3.1. Naïve Selection

The naïve selection approach, used as baseline in the RATS SAD evaluation, is given by the following simple algorithm: (1) given a certain random seed, sort the available files from the channel of interest; (2) select audio starting from the start of the first file going down the list until collecting B minutes of audio. In this method, all selected files except the last one are used fully for annotation.

3.2. Passive Selection

Another baseline used in the RATS SAD evaluation is given by the following algorithm: (1) sort files as for the naïve baseline; (2) select audio from top to bottom files restricting selection to the regions hypothesized by a baseline (unadapted) system as being speech, after padding each speech region with two seconds on each side, until B minutes of audio are collected. This system simulates a scenario where the speech regions detected by a baseline (unadapted) system are given to a user who corrects the output when errors are found. The corrected detections can then be used as annotations for adaptation. The padding of speech regions is done to simulate the process a user might use of listening for a moment before and after each hypothesized region. The regions selected by this approach will only include misses incurred by the baseline system when they occur within two seconds of a detection speech region.

3.3. High Coverage Selection

Our proposed approach for region selection is as follows: (1) split the requested duration B evenly across all available files;

(2) if the resulting duration per file is less than a certain minimum duration M, randomly choose N files such that each file gets at least this minimum duration assigned to it; (3) within each file, decide the location of the requested audio using different criteria. In turn, the criteria for locating regions within each file can be: **evenly-spread**, where the assigned duration is divided into snippets of duration M, and the snippets are evenly spread across the file; and **uniform**, where snippets of duration M are located to sample the LLRs for that signal as uniformly as possible. This last option requires, as the passive selection approach, that a baseline system is run to obtain the LLRs for each signal. Once LLRs are computed and smoothed, as described in Section 2, the mean LLRs over snippets of length M shifted one frame at a time are computed. A value is then randomly chosen from a uniform distribution within the minimum and maximum of these LLR means. The M-duration snippet with the mean LLR closest to this value is then selected. Next, another value is drawn from the uniform distribution and the snippet with the closest LLR is selected. If this region overlaps with the already selected region, its borders are expanded around that region to achieve a total duration of 2M, including the old snippet and the new one (starting by expanding toward the end and reversing toward the beginning of the file only if not enough frames are available toward the end). This process continues until all snippets are allocated. This way, when the assigned duration for a certain file is significantly smaller than its total length, the distribution of the mean LLR of the selected snippets will be close to a uniform distribution. Once the assigned duration for a file approaches its length, the distribution of the mean LLR of the selected snippets will approach the actual LLR distribution of the file, since almost the full file will be selected.

In all cases, we restrict a certain proportion (given by a tunable parameter) of the files to have one region located at the introductory part of the file, defined to be the first 15 seconds of each file. The purpose of this restriction is to ensure that any special effects present at the beginning of the file, like handshaking events between transmitter and receivers, hold-music when calling a certain phone number, answering machine message, etc., are well represented in the selected regions.

4. Multi-Condition SAD Database

The training data and the test data used in the experiments came from to the Linguistic Data Consortium (LDC) collections for the DARPA RATS program [7]. Conversational telephone recordings (called source signals) were retransmitted using a multilink transmission system at LDC. Several combinations of transmitters and receivers were used to retransmit, resulting in extremely noisy and distorted signals. For training, we used the RATS SAD training data. Channel D was not included in the SAD data due to annotation problems. The clean source was included among the channels. This data contained 830 hours of speech and 677 hours of non-speech. Given the large size of this data set, during DNN training, we selected 1 every 10 frames.

For testing, we used four different types of channels, some created by corrupting source signals from the RATS dataset. Table 1 shows statistics on the different channel groups.

RATS: This group of channels is extracted from the RATS novel channel collection, released by LDC in 2014. This data was created by using different transmitter/receiver pairs, new transmitter/receiver locations and longer distances than those used in the original set used for model training in this work. We used eight of the released channels (A, D, G, H, K, M, Q, and R), discarding the ones with clear annotation problems and omitting some channels for future evaluation on a held-out set (these channels are not used for the results in this paper). During development, the adaptation and test signals from those channels with gross annotation errors were discarded.

Table 1: Statistics for different channel groups: number of channels, average adaptation size (in minutes), average test size (in minutes) and average proportion of speech content on the test data. Averages are taken across channels within the group.

Group	#Chan	Adapt	Test	Prop. Speech
RATS	8	96.9	54.0	38
RATS LSD	8	68.8	36.5	8
Codec	13	108.6	114.1	41
Music	4	108.6	114.1	41

RATS LSD: We created low speech-density (LSD) data starting from the RATS unseen channels above. To this end, for each signal, a portion of the speech regions was randomly selected for retention, adding up to a maximum of X% of the total final file length, where X was chosen from a uniform distribution between 0 and 10. The speech regions not selected for retention were cut from the audio. The resulting speech and non-speech in the signal had characteristics identical to those of the original signal. The goal of creating this data was to test whether algorithms behave differently depending on the proportion of speech present in the adaptation data. As we will see, this is definitely the case, at least for some algorithms.

Codec: We created transcoded signals starting from 58 source signals used for retransmitting the RATS data. Each of these signals was encoded with a few different encoders and transcoded back to sphere format. The encoders included in the results in this paper are: AMRNB (12.2KHz and 5.9KHz), CODEC2 (2.4KHz), G723-1 (6.3KHz), OPUS (4KHz and 8KHz) and OPUS-VBR (4KHz and 8KHz). The numbers in parenthesis correspond to the encoding rates considered. For four of these encoders we transcoded the signals six times in a row to obtain, in some cases, highly degraded signals.

Music: We created signals corrupted with non-vocal music at different signal-to-noise ratios (SNRs). We added different short snippets of non-vocal music to the same source signals used to create the Codec data. The music types were classic, jazz and modern.

5. Experimental Setup

We trained a DNN SAD system by using the training data described in Section 4. The DNN took 13-dimensional MFCC features, normalized to have mean 0 and standard deviation of 1 over each file and each dimension. The DNN contained two hidden layers of sizes 500 and 100. This baseline DNN was used to obtain the LLRs required for some of the selection approaches and also used as the initial model for adaptation. The different approaches described in Section 3 were used to select regions for which annotations were retrieved, simulating the process of annotation that would occur in practice. The features for those regions as well as the annotations were then used to adapt the baseline model to the channel of interest.

To analyze the effect of the different selection algorithms as a function of the annotation time B, we swiped B logarithmically, from 1 to 32 minutes. We also show results for full adaptation, where all available adaptation data for each channel was used for adaptation. Each selection approach was run 10 times with different random seeds and performance and the time was averaged across runs for each value of B. For the full-

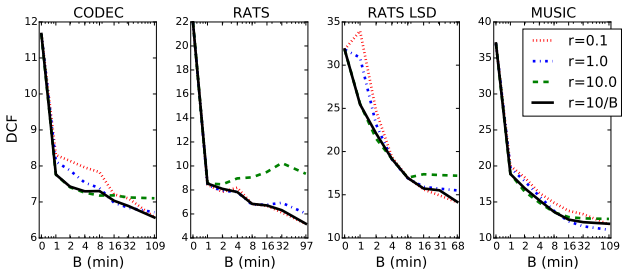

Figure 1: Results when setting the regularization parameter r to different fixed values and to 10/B. B, the x-axis, is the amount of time annotated for adaptation (in minutes).

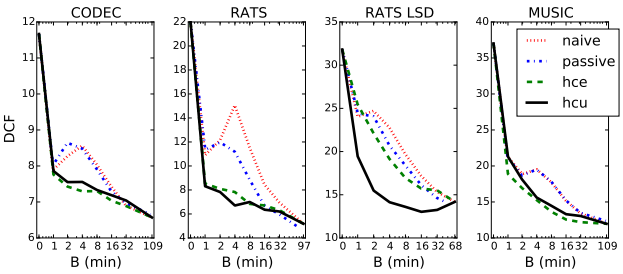

Figure 2: Results comparing different selection approaches: naïve, passive, high coverage with evenly-spread region location (hce) and high coverage with uniform region location (hcu). The x-axis, B, is the same as in Figure 1.

adaptation case, the seed was still used to shuffle the adaptation features, resulting in 10 different results, although the data was identical in all cases.

Two types of error can be computed for SAD: (1) the miss rate (the proportion of speech frames labeled as non-speech) and (2) the false alarm rate (the proportion of non-speech frames labeled as speech). In Phase 4 of the RATS program, a "forgiveness" collar of 2.0 seconds was used around all annotated speech regions. False alarm errors over those regions were disregarded. We use this same collar for our results. As in our previous SAD paper [4], here we used the actual DCF (or just DCF for brevity) as the metric. We do not present the minimum DCF values for lack of space. To obtain the DCFs we post-processed the LLRs from each of the systems as described in Section 2 by using an average filter of 41 frames, thresholded them at 0.0, padded each resulting speech region with 0.3 seconds on each side, and finally summed false alarm and miss rates for each channel to calculate the DCF. The reported results are averages across channels and random seeds within groups of channels.

6. Results

Figure 1 shows the results when using different fixed values for the regularization parameter (0.1, 1.0 and 10.0) and when using the proposed approach, where the parameter was set to 10/B (B being the total annotated time in minutes). These results correspond to the high coverage (HC) selection approach, with evenly spread snippets of two second duration, forcing 50% of the files to have one snippet located at the beginning of the file. Clearly, fixed values of the regularization parameter cannot accommodate all possible values of B: large values of regularization are good for small B, while small values are good for large B. The proposed approach, setting this parameter to 10/B, strikes a good trade off, and is optimal in most cases. This is the approach we used for setting the regularization value for subsequent results.

Figure 2 shows a comparison between the baseline selection approaches and the two proposed ones. We see that, among the two baseline approaches, the passive one is slightly better for the two groups of RATS channels. This finding is most likely due to the fact that the passive approach selects regions from more unique files than the naïve approach, because the naïve approach uses the full files to reach the budget, while the passive one only uses padded speech regions. Hence, the passive approach indirectly enables for more coverage of files than the naïve one for the same B.

The two proposed high coverage approaches are quite comparable when the proportion of speech is close to 50% (RATS, Codec, and Music). When this is not the case, for the RATS LSD group, the uniform selection of regions gives a very clear advantage over the evenly spread one. This is because the evenly-spread selection results in a low proportion of speech frames being selected for annotation (and, hence, for adaptation), approximately the same proportion available in the full signal. On the other hand, the uniform selection criteria results in a higher percent of speech regions being selected, which clearly benefits the final adapted model. For example, the percent of speech in the selected regions for B=16 is 18%, versus 8% for all available data. We believe this is the reason why, for this group of channels, selecting only 16 minutes leads to better results than using all of the available data.

Finally, we note that our results indicate that, as long as the regularization factor is chosen as proposed, the performance on seen channels (that is, the channels present in the training data for the baseline system) is not significantly affected by adapting to data obtained with any of the presented selection techniques. These results are not shown due to lack of space.

7. Conclusions

In this work, we take the first steps towards the application of active learning ideas for the SAD task. Our results demonstrate that simple techniques can greatly improve performance over a naïve approach consisting on selecting N full files adding up to a desired budgeted duration. We also show that an approach that uses the frame-level scores from a baseline system to select regions such that the score distribution is uniformly sampled can lead to significant gains when the signals only contain a small proportion of speech. Further, we show that careful selection of the regularization parameters during adaptation is essential. Finally, the results indicate that, for some channels, just a few minutes of carefully selected data can lead to results comparable to those obtained with one to two hours of adaptation data. In the future, we plan to continue working on selection techniques, including options that preferentially select regions from certain signals that might result in more useful adaptation. We will also explore algorithms that attempt to select regions that are purely speech or non-speech, which would greatly simplify the annotation process.

8. Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0037. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Distribution Statement A: Approved for Public Release, Distribution Unlimited.

9. References

[1] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on YouTube using deep neural networks." in *Proc. Interspeech*, Lyon, France, Aug. 2013.

[2] J. Ma, "Improving the speech activity detection for the DARPA RATS phase-3 evaluation," in *Proc. Interspeech*, Singapore, Sep. 2014.

[3] S. Thomas, G. Saon, M. Van Segbroeck, and S. S. Narayanan, "Improvements to the IBM speech activity detection system for the DARPA RATS program," in *Proc. ICASSP*, Brisbane, Australia, May 2015.

[4] L. Ferrer, M. Graciarena, and V. Mitra, "A phonetically aware system for speech activity detection," in *Proc. ICASSP*, Shanghai, China, March 2016.

[5] B. Settles, "Active learning literature survey," *University of Wisconsin, Madison*, vol. 52, no. 55-66, p. 11, 2010.

[6] G. Gelly and J.-L. Gauvain, "Minimum word error training of rnn-based voice activity detection," in *Proc. Interspeech*, Dresden, Sep. 2015.

[7] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Odyssey 2012: The Speaker and Language Recognition Workshop*, 2012.