Weakly Supervised Training of Noisemes

Tianyu Xu
Carnegie Mellon University
School of Computer Science
Language Technologies Institute
Email: tianyux@andrew.cmu.edu

Liping Xiong
Carnegie Mellon University
School of Computer Science
Language Technologies Institute
Email: lipingx@andrew.cmu.edu

Abstract

In this paper, we present a novel approach to enlarge noiseme training dataset inspired by the idea of co-training. Given videos with automatically generated noiseme labels and image concepts, we introduce three different methods to discover stable correlations between sounds and images, and use corresponding image concepts as the evidence to support the automatically labeled noisemes based on insightful correlations. We propose several ways to fuse and refine the found correlations. Also, we construct new training datasets in two different ways, and conduct several experiments to show that the proposed approach is able to effectively select new training samples from automatically labeled data, and weakly supervised training can help improve the accuracy of noiseme detectors.

Keywords

noiseme, co-training, multimedia analysis.

I. Introduction

Noiseme, as described in [1], is the smallest segmental unit of sound employed to form meaningful contrasts between noises. Noiseme labels describe distinct noise units based on audio concepts, independent of visual concepts as much as possible. They are useful for many multimedia analysis tasks such as multimedia event detection and detailed summaries of videos. We can create event signatures or event fingerprints using particular noiseme patterns or significant noiseme co-occurrences from different modalities.

In order to build accurate noiseme detectors, a large-scale manually labeled training set is needed. However, current noiseme classifiers were only trained on 5.6 hours of manually labeled audio data, which is not sufficient and leaves much room for performance improvements. We want to find a way to leverage the 6,000 hours videos we have now, although they do not have manually labeled noiseme information, and let them contribute to the training process as well. To achieve this goal, we can use the noiseme detectors trained by manually labeled data to label the 6,000 hours unlabeled data, and select reliable noiseme snippets as new samples to enrich the training set.

We hold the assumption that some noisemes can be ”seen” in the corresponding video. For example, if we see a dog appears in a video, we are likely to hear barking at the same time. Similarly, if we see a vehicle drives through the street, we will probably hear the engine sound fade in and fade out. Holding the assumption that some noisemes are correlated with some image concepts, we can find informative noiseme-image correlation pairs and use the image concepts as the evidence to support automatically generated noiseme labels. Then we select verified noiseme features from automatically labeled data to enlarge the training set and re-train the models.

We propose the weakly supervised training approach as four phases. First, we describe three different correlation discovering methods which can find correlation pairs, i.e. find noiseme labels and image concepts that co-occur frequently. Second, we introduce a straightforward late fusion process to combine correlation discovering results. Third, we refine the resultant correlations from two different aspects: eliminate inaccurate automatically generated labels and take advantage of image concept hierarchy to merge similar image concepts. Fourth, we describe how to construct a new training set based on the correlations we get.

The rest of the paper is organized as follows. In Section II, we give an overview of the datasets we have, including the manually labeled set and the much larger set with automatically generated labels, and the low-level acoustic features we use in noiseme classifier training. In Section III, we provide details of the proposed approach and show how to find informative correlations and build new training set based on them. In Section IV and Section V, we discuss the experiments we conducted and analyze the results. At last, we give our conclusions and indicate who contributed which part of the project.

II. Data and Features

We have two different datasets. The first dataset is the one manually labeled with noiseme labels, and its duration is 5.6 hours. The human ear is fairly good in identifying sounds so we can take these manually labeled noisemes as ground truth. The second dataset is the one with automatically generated noiseme labels and image concepts. The total duration of this dataset is more than 6,000 hours. We have 1,000 different image concepts, which were labeled by a koala model. For each video, several keyframes were selected and labeled with these image concepts. Also, we have 40 different noiseme labels, which were labeled by a random forest model. Note that the number of noiseme labels is not consistent with the number we provided in Section I, because some noisemes only appear in several clips so that we filtered them out. A fixed length (100ms) sliding window slides from the beginning to the end and labels are given by the noiseme detectors trained on the manually labeled dataset. The format of each video is shown in Figure 1.

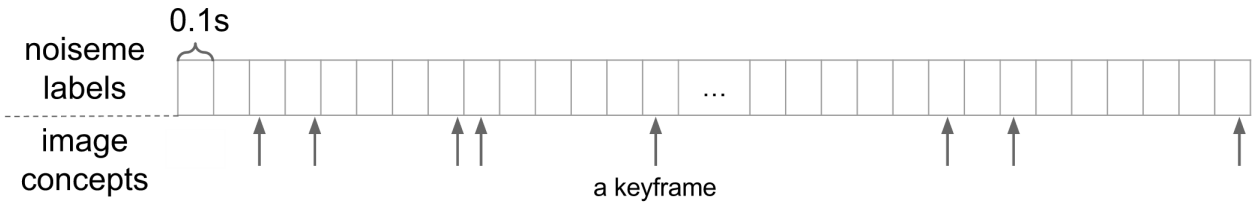

Fig. 1: The format of videos in the automatically labeled dataset.

We use features described in [2] as low-level features. First, standard acoustic features are extracted using the toolkit openSMILE [3] (speech and music interpretation by large space extraction), and then we apply a set of statistical functions over these acoustic features within the fixed length sliding window to get the final representations. The dimension of the feature vector is 983. The feature set is listed in Table I.

TABLE I: Feature set list, cited from [2].

MFCC (13 + 13 delta + 13 delta delta)	arithmetic mean, standard deviation, linear regression parameter, quartiles, range
MFC (26 + 26 delta)	The same set with log-energy + kurtosis
Log-energy	quadratic mean (non-zero value), standard derivation, max Pos, min Pos, linear regression parameter, quartiles, range
F0	quadratic, standard deviation, maximum mean, minimum mean, linear regression, qiartiles, range

III. Methods

A. Correlation Discovering

We proposed three different approaches to discover the insightful correlation pairs, i.e. find noiseme labels and image concepts that co-occur frequently. All the approaches are based on statistical models and no prior knowledge is required toward the correlations.

1) Score multiplication: The first approach is the most straightforward approach. Automatically generated noiseme labels and automatically generated image concepts are both assigned with confident scores range from $[0, 1]$. For each noiseme - image concept pair, we can simply multiply their confident scores as the *correlation score* towards the correlation pair, since we expect a good correlation if both scores are high. If either of the scores is low, then their is no evidence can show that the image concept co-occur with the noiseme label. To discover the potential interesting pairs, for each video keyframe, we find the corresponding audio window, and then for each image concept in the keyframe and each noiseme label in the window, we multiply their confidence scores.

2) Conditional Probability: The second approach is based on conditional probability theory. The assumption is intuitive: if a noiseme and an image concept co-occur a lot, then given the prior knowledge that an image concept occurred, the probability of the corresponding noiseme labels should be higher compared to the case without the prior knowledge. So we compute

$$P(\text{noiseme}) = \frac{\text{\# of windows with noiseme}}{\text{\# of windows}} \tag{1}$$

$$P(\text{noiseme}|\text{image}) = \frac{\text{\# of keyframes with noiseme and image}}{\text{\# of keyframes with image}} \tag{2}$$

and take the ratio of Equation 2 and Equation 1 as the correlation score of given noiseme label and image concept.

3) Binomial Ratio Likelihood Test: The third approach is based on *Binomial Ratio Likelihood Test* (BRLT). Given an image concept c and a noiseme n, consider two different test:

1) Draw keyframes with noiseme label n from all the keyframes with image concept label c.
2) Draw keyframes with noiseme label n from all the keyframes with image concept labels other than c.

Suppose we draw n_1 times and got k_1 success in test 1, draw n_2 times and got k_2 success in test 2. If there is a correlation between n and c, then these two tests should differ a lot. We use BRLT to check if the test results are from one binomial (i.e. k_1/n_1 and k_2/n_2 were different due to chance) or from two distinct binomials.

To compute the BRLT score, we use the following equations:

$$BRLT(k_1, n_1, k_2, n_2) = 2 \times \log\left(L(p_1, k_1, n_1) \times \frac{L(p_2, k_2, n_2)}{L(p, k_1, n_1) * L(p, k_2, n_2)}\right) \tag{3}$$

where

$$L(p, k, n) = p^k \times (1-p)^{n-k} \tag{4}$$

$$p = \frac{k_1 + k_2}{n_1 + n_2} \tag{5}$$

$$p_i = \frac{k_i}{n_i} \tag{6}$$

B. Correlation Fusion

To combine the discovered correlations from three different approaches, we use late fusion to combine the results. We merge all the correlation pairs from three sources, and take the weighted sum of the scores as the final score. We tried different weights, and finally the three approaches are weighted as 0.2, 0.2, 0.6 respectively, because we got a relatively good correlation result based on this setting. However, the weights are actually not that important, since they can only change the ranking of the correlation list.

C. Correlation Refinement

We refine the correlation list from two different aspects.

First, some of the noiseme labels and image concepts are inaccurate, thus the correlations based on them are not reliable and must be removed. Otherwise, the errors from imperfect noiseme classifier and imperfect image concept detectors will accumulate and impact the result of our final performance. The total duration of training samples is a good estimation of the corresponding noiseme classifier's accuracy, since noiseme classifiers trained with larger dataset will be more accurate. For image concepts, we use a pre-existing test result of the image labeling accuracy, where each image detector was tested with ten samples. We remove noisemes whose classifiers were trained with insufficient data, and remove image concepts whose classifiers have accuracy less than a certain threshold.

Second, we can leverage the image concept hierarchy information to merge similar image concepts. For example, the noiseme "animal-dog" can be associated with both German Terrier and Chiwawa. If we take advantage of image concept hierarchy and consider these two different dog breeds as a more general concept "dog", we can further more polish the ranking of the correlation list.

D. Training Set Construction

After performing all the steps we described in the previous sections, we get a informative correlation list, in which each pair of noiseme and image concepts co-occur frequently. We then use this correlation information to construct new (and much larger) training set.

We first read the 6,000 dataset again, and for each keyframe, we check the corresponding audio window and see if the noiseme-image pair is in the *target correlation* set. If the pair exists in the list, then the keyframe is a *target keyframe*. We can then select acoustic features of related audio windows with regard to the target keyframe. There are two different kinds of related audio windows. First, the audio window which target keyframe lies in is a related window. Second, the contiguous windows with the same noiseme label are related windows. We combine the features of these related windows and add proper noiseme labels on them to build a new training set.

IV. Experiments

The experiments were conducted on 5.6 hours of manually labeled data and 6,000 hours of automatically generated data. The baseline experiment used only manually labeled data for training and testing. Experiment 1 examined the accuracy of the models trained only on the automatically generated data. Experiment 2 combined the manually labeled data and automatically generated data. Experiment 3 inspected the effect of the number of correlations on the performance of the models. Based on the results of the previous experiments, in experiment 4, we examined the effect of image concept accuracy on the performance of the classifiers because we found that the quality of the correlations found using 30%+ image concept accuracy is not very good. And we further designed experiment 5 and 6 to test our conjecture that it is the more accurate correlations that contribute to the performance improvement of the detectors.

Each experiment used different training set, but they were all measured on 40% manually labeled data, which is selected randomly for each experiment. we applied the following strategy in randomly selecting training or test samples. Instead of randomly selecting complete videos from all the videos, we randomly selected snippets from all of the video snippets.

For measurement, we used the `precision_recall_fscore_support()` method from *sklearn*, which can give back the precious, recall and F1-score of the model. Each time we changed the training set, we ran the training and test 10 times and took the mean precision, mean recall and mean F1-score of the 10 times running as measurements.

In these experiments, we inspected two different definitions of related audio windows as illustrated in Section III-D. The first one is selecting the following windows of length 1s for each target keyframe. The second one is select contiguous windows of length 5s with the same noiseme label for each target keyframe.

Out of 40 noisemes, we filtered out those occurred rarely according to [2], and finally we got 22 noisemes. All our experiments were conducted on those 22 noisemes.

V. Results and Discussion

Our baseline is the classifiers trained using randomly selected 60% of manually labeled data, and we tested on 40% of manually labeled data. The result is shown in Figure 2.

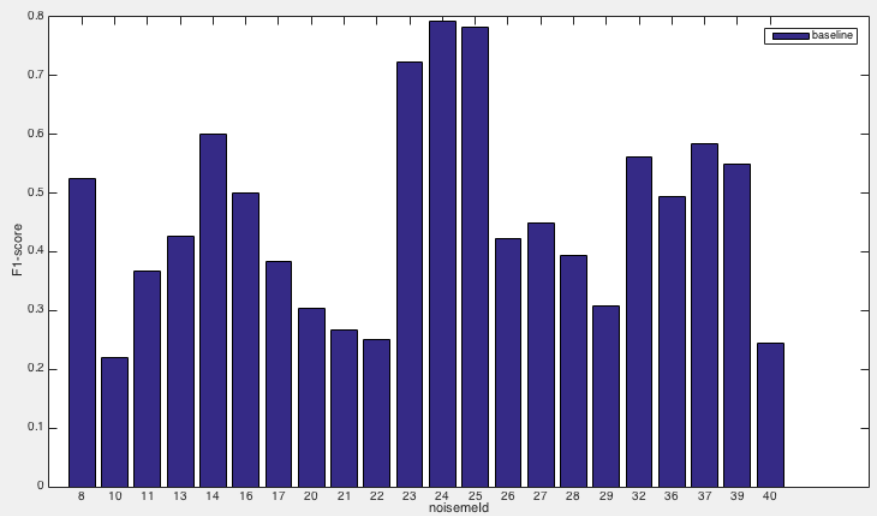

Fig. 2: The mean F1-score of the baseline.

Then we conducted several experiments in which we tried to beat the baseline.

A. Experiment 1

In experiment 1, we used only the automatically generated data to train the classifiers, and we tested on the 40% manually labeled data. The result is shown in Figure 3. As we can see, the F1-scores of these classifiers are far lower than the baseline.

We think there are two reasons for it. The first one could be that the low accuracy of the training set constructed from the automatically labeled data leads to the bad performance of the classifiers trained on them. The second reason may be that the baseline classifiers were trained and tested on the same set of data, while the new classifiers were trained and tested on two different sets of data (trained on automatically labeled data and tested on manually labeled data). These two sets of data can be very different.

B. Experiment 2

In experiment 2, we combined the manually labeled data and the automatically generated data. 60% of manually labeled data was randomly selected and combined with automatically generated data (using all of the 5167 correlations) to train the classifiers. The result is shown in Figure 4. The result is almost the same as the baseline, which indicates that the training set constructed from automatically generated data does not help to improve the accuracy. The reason may be that the new training set we built has too much noises because we used all the found correlations. The correlations with low scores may introduce noises into training data.

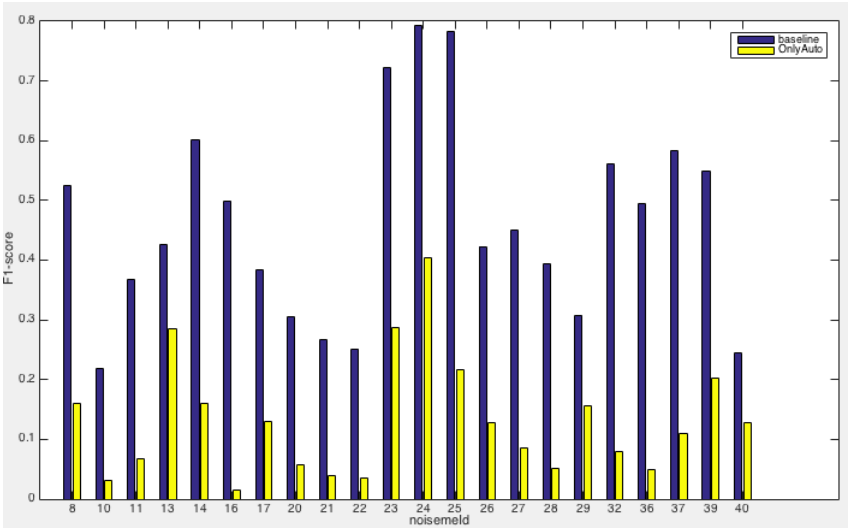

Fig. 3: The mean F1-score the baseline and classifiers trained using only automatically generated data.

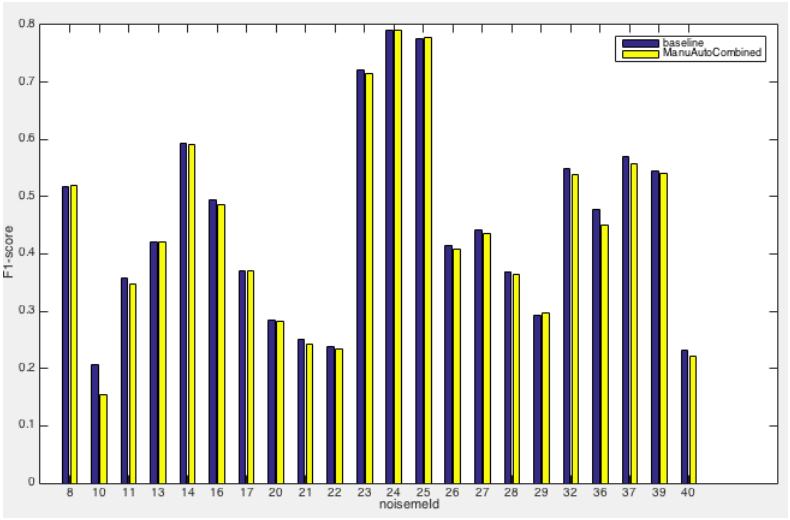

Fig. 4: The mean F1-score of the baseline and classifiers trained using both manually labeled and automatically generated data.

C. Experiment 3

Therefore, in experiment 3, we inspected the effect of different sets of correlations on the performance of the classifiers. In each experiment, top 5000, 3000 and 2000 correlations were selected. The result is showed in Figure 5.

There is not much difference between these experiments. We guess the reason may be that even the high ranking correlations have low accuracy (since when we looked over the top 50 correlations, there were still some correlations that do not make sense by intuition). If the image concepts are not correct, then there is no way that the correlations will be in high quality. So we designed the next experiment to test our conjecture.

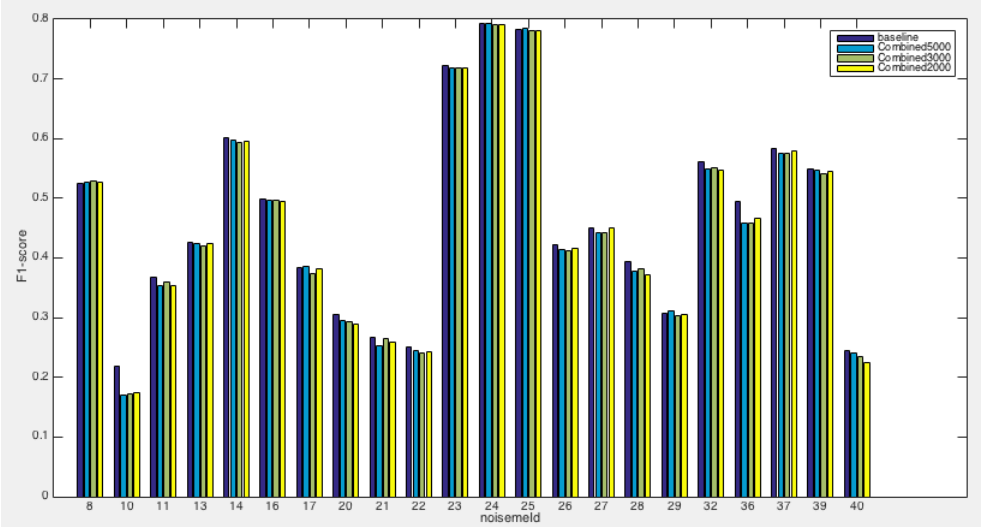

Fig. 5: The mean F1-score of baseline and classifiers trained using different amount of correlations.

D. Experiment 4

In experiments 4, we tried to improve the quality of the correlations by increasing the accuracy threshold of image concepts. Only image concepts with accuracy 80%+ were selected, which led to 1,256 correlations. And, since there are less correlations, in order to get a comparable amount of training data as previous experiments, we changed the way of selecting training set to selecting contiguous windows of length 5s with the same noiseme label (the second definition of related audio windows as illustrated in Section III-D). The result is shown in Figure 6.

We found that 15 out of 22 classifiers were improved. The maximum increment of F1 score is 0.025 (9%). The average increment of F1 score is 0.01 (4%).

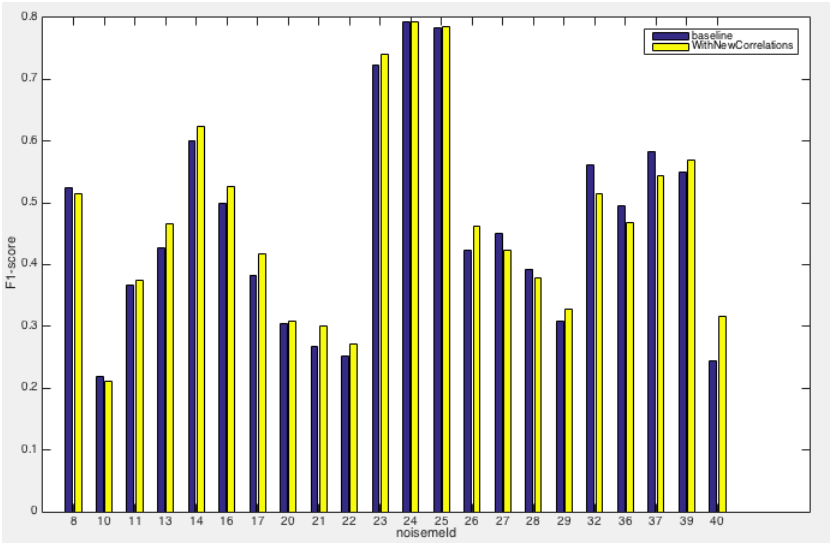

Fig. 6: The mean F1-score of the baseline and the classifier trained using new set of correlations.

After proving the effectiveness of the methods in this experiment, we tried to figure out whether it is because of the less but more accurate correlations or the usage of different definitions of related audio windows.

E. Experiment 5

In experiment 5, we used the top 1,265 correlations got from image concepts with accuracy greater than 30%. We tried to compare it with using all of the 1,265 correlations obtained from image concepts with accuracy greater than 80%. And we were using the second definition of related audio windows in both experiments, we could compare the effect of the two sets of correlations. The result is shown in Figure 7. 17 out of 22 classifiers are improved, and the average increment of F1 score is 0.009. This is a proof that the new correlations did contribute to the performance improvement of the detectors.

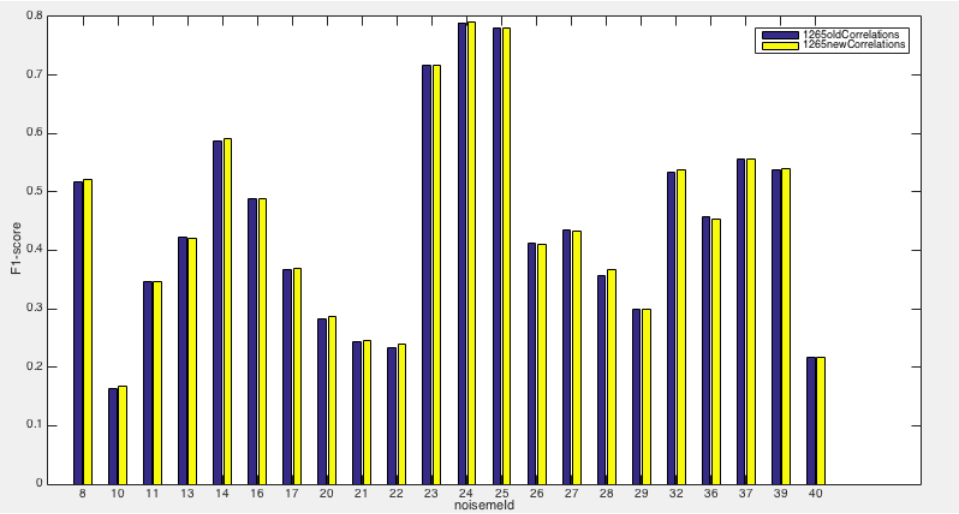

Fig. 7: The mean F1-score of classifiers trained using different set of correlations.

We also tried to use all of the 5,167 correlations and the second definition to construct the training set ,But we found that the set is too big to be used by our current method.And from experiment 6,we can see that the definition of related windows actually does not matter so much.

F. Experiment 6

In this experiment we tried to test whether the second definition of related audio windows is better or not. So we keep the correlations the same, and used the first and second definition respectively.The result is shown in figure 8. We found that 11 out of 22 classifiers were improved.The average increase of F1 score is 0.0025. However, 11 out of 22 classifiers degraded.The average decrease of F1 score is 0.0028. Therefore, there is no much difference between the two definitions. So, this is another proof that the improvement in experiment 4 is due to the high quality of the correlations, not the new definition of related audio windows.

Apart from all of the above experiments, we also conducted lots of other experiments, such as cross validation on the manually labeled data set and the combined data set. But the above ones are those we think have the most information.

VI. Conclusion

As shown in the previous experiments, we come to the conclusion that we can found informative noiseme-image label correlations based on statistical approaches. And with the benefit of these correlations, we can use an image label as an evidence of an automatically generated noiseme label, select those sound snippets as new training data to enrich the training dataset, and get better noiseme detectors. Also, we get the conclusion that the quality of discovered correlations matters and the image concept accuracy is a bottleneck.

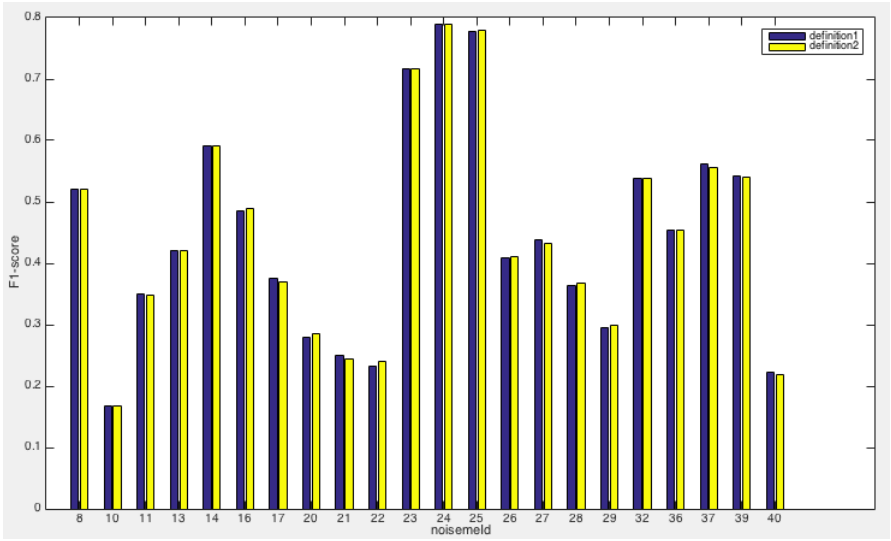

Fig. 8: The mean F1-score of classifiers trained using different definition of related audio windows.

VII. Group Contribution

Tianyu Xu proposed and implemented all the algorithms related to noiseme-image correlations, i.e. correlation discovering, correlation fusion, correlation refinement, and new training set construction. Liping Xiong designed and conducted all the experiments, and gave detailed analysis on them. As for the report, Tianyu composed Abstract, Introduction, Data and Features, Methods; Liping is in charge of Experiments, Results and Discussion, Conclusion. Overall, Tianyu and Liping have similar workload in this project.

Acknowledgment

We would like to thank Professor Florian Metze for his detailed and patient guidance during this semester. We also want to thank Shoou-I Yu, who provides the accuracy information of image concepts, and Shicheng Xu, who provides the image concept hierarchy data.

References

[1] S. Burger, Q. Jin, P. F. Schulam, and F. Metze, "Noisemes: Manual annotation of environmental noise in audio streams," *submission to Interspeech*, 2012.

[2] Y. Wang, S. Rawat, and F. Metze, "Exploring audio semantic concepts for event-based video retrieval," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1360–1364.

[3] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.