Data Selection for Acoustic Emotion Recognition: Analyzing and Comparing Utterance and Sub-Utterance Selection Strategies

Duc Le, Emily Mower Provost
Computer Science and Engineering
University of Michigan, Ann Arbor, MI, USA
{ducle, emilykmp}@umich.edu

***Abstract*—Data selection is an important component of cross-corpus training and semi-supervised/active learning. However, its effect on acoustic emotion recognition is still not well understood. In this work, we perform an in-depth exploration of various data selection strategies for emotion classification from speech using classifier agreement as the selection metric. Our methods span both the traditional utterance as well as the less explored sub-utterance level. A median unweighted average recall of 70.68%, comparable to the winner of the 2009 INTERSPEECH Emotion Challenge, was achieved on the FAU Aibo 2-class problem using less than 50% of the training data. Our results indicate that sub-utterance selection leads to slightly faster convergence and significantly more stable learning. In addition, diversifying instances in terms of classifier agreement produces a faster learning rate, whereas selecting those near the median results in higher stability. We show that the selected data instances can be explained intuitively based on their acoustic properties and position within an utterance. Our work helps provide a deeper understanding of the strengths, weaknesses, and trade-offs of different data selection strategies for speech emotion recognition.**

***Keywords*—speech emotion recognition; data selection; dynamic modeling; artificial neural network; classifier agreement**

I. Introduction

Emotion recognition from speech is an important problem for many domains, including human-computer interaction and clinical applications. A major challenge facing this research area is that there exists a heterogeneous set of manually collected and labeled datasets. Two important problems need to be addressed to widen the horizon of current research: (1) to develop training methods that can learn effectively from heterogeneous data sources and generalize well across different emotion corpora, and (2) to develop techniques that can leverage the large amount of unlabeled speech data in an active learning or semi-supervised framework.

There have been a number of works that have examined these two problems, which we summarize in Section II. Fundamental to both problems is the key issue of selecting training data in a way that facilitates accurate and stable emotion classification in terms of within- and across-run variation. While previous works have proposed various methods for data selection, their effect on classification performance is still not well understood. In addition, there has been limited investigation into the characteristics of selected data instances. Having a deep understanding of the strengths, weaknesses, and trade-offs of different data selection strategies will be an important step toward creating large-scale emotion recognition systems that can operate in real-world settings.

In this paper, we perform an in-depth investigation of various data selection methods and their impact on classification performance in the 2-class sub-challenge of FAU Aibo, a benchmark dataset in acoustic emotion recognition [1]. We explore selection methods at both the traditional utterance and less common sub-utterance level, where only a portion of an utterance is selected for training. Sub-utterance data selection is motivated by the hypothesis that not all parts of an utterance are equally indicative of the target emotion, thus identifying and training only on salient local regions will facilitate generalization. Previous works have demonstrated the importance of local emotion dynamics in both human perception [2] and automatic classification [3]–[6].

We use classifier agreement, combined with four different ranking methods, to order data instances for selection. In Section VI-A, we show that diversifying training data in terms of classifier agreement leads to faster learning, achieving a median unweighted average recall of 70.68% with less than 50% of the training data. This result is comparable to the performance of the full training set as well as the winner of the 2009 INTERSPEECH Emotion Challenge [7]. In contrast, selecting instances with medium agreement results in higher stability. Sub-utterance selection produces a slightly better convergence rate than utterance selection and has significantly less variation within individual and across multiple runs. Our analysis of selected data instances in Section VI-B demonstrates that classifier agreement is correlated with pitch, intensity, and a segment's relative position within the utterance. The contributions of our work lie in the comprehensive review of various selection methods, the novel sub-utterance data selection, and the in-depth analysis of selected data samples.

II. Related Work

Data selection strategies for acoustic emotion recognition have been investigated directly in various works. Erdem et al. [8] and Bozkurt et al. [9] applied data selection using the well-known Random Sample Consensus (RANSAC) algorithm [10], which searches for a random subset of utterances that induces a model having the highest consensus with the training

set. They found that RANSAC-based training can significantly improve the classification performance of 1-state Hidden Markov Model (HMM). Schuller et al. proposed selecting utterances in decreasing level of prototypicality, defined as an instance's distance to its class center in the feature space [11]. They achieved good performance in cross-corpus arousal recognition when using only 50% of the training data. Meudt and Schwenker showed that discarding misclassified utterances and selecting the most confidently recognized instances among the remaining data can improve test accuracy and reduce standard deviation [12]. Zhang et al. found that classification performance is improved by removing utterances with low human agreement from each emotion class, followed by random subsampling for class balancing [13].

Data selection for speech emotion recognition has also been studied indirectly as a component of an active learning or semi-supervised system. In this research paradigm, ground-truth labels for training data are not available beforehand, and the system must select subsets of data for machine or human annotation. The works by Zhang et al. focused on active learning frameworks where utterances are selected based on class sparseness, medium classifier confidence, or medium predicted labeler uncertainty [14], [15]. They found that selecting based on class sparseness outperforms the other two methods without instance upsampling, while the latter strategies give good results after applying upsampling. Mariooryad et al. showed that emotional utterances can be identified from a neutral speech corpus using the confidence value of a Support Vector Machine (SVM) trained on emotional speech [16]. Zhang et al. combined active learning and semi-supervised training by using the most confidently predicted utterances for machine annotation and passing those with medium confidence onto human labelers [17]. Their system was able to achieve good performance with relatively few human-labeled instances.

Our paper differs from these works in three aspects. Firstly, we provide a comprehensive analysis and comparison of the effect of different data selection strategies on classification performance, namely the rate and stability of learning. This type of detailed comparison has been under-explored in previous works, thus the strengths, weaknesses, and trade-offs of different selection methods are still not well understood. Secondly, we propose and analyze novel sub-utterance data selection strategies which have not been considered in these works. Thirdly, we investigate the characteristics of selected data instances to provide an intuitive understanding regarding the type of data most useful for classification. This set of analysis has not been performed in previous works.

III. FAU Aibo Emotion Corpus

FAU Aibo [1] is a benchmark dataset used in the 2009 INTERSPEECH Emotion Challenge [7]. It contains recordings of 51 German children interacting spontaneously with the pet robot Aibo. The recordings took place in two different schools, Ohm and Mont, which formed the training and test set, respectively. Speech recordings were manually segmented into syntactically and semantically relevant chunks which convey relatively homogeneous emotions and were shown to be the best unit for automatic analysis. The dataset was initially labeled at the word level by five human annotators. These labels were heuristically aggregated into chunk-level ground-truths. The training set includes 26 speakers and 9,959 chunks. The test set consists of 25 speaker and 8,257 chunks. For consistency, we henceforth refer to chunks as utterances.

In this work we tackle the 2-class sub-challenge of FAU Aibo. This setup makes it easy to compare our results with previous work. Each utterance is assigned one of two labels: **IDL**, which represents a quasi-neutral state, and **NEG**, which exemplifies negative valence [1]. In both the training and test set, the ratio of IDL to NEG is roughly 2 to 1. Because of this imbalance, a classifier's performance is evaluated using unweighted average recall (UAR), defined as the mean per-class accuracy. The baseline UAR, 67.7%, was achieved by static modeling with linear-kernel SVM, instance upsampling, and standardization [7]. The winner of the 2009 challenge performed logistic regression fusion of three systems using short- and long-term speech features, achieving a UAR of 70.3% [18]. Details about the challenge and other participants can be found in [19]. The current state-of-the-art UAR on this problem is 72.8%, achieved by compensating for channel and speaker differences across the training and test set [20].

IV. Methods

A. Hybrid HMM-ANN Emotion Classifier

The classifier we use in this paper is based on our previous work, which achieved state-of-the-art results on FAU Aibo's 5-class problem [6]. In this earlier work, we modeled each emotion dynamically as a HMM where the emission probabilities are estimated using a 5-layer Deep Belief Network (DBN) [21] trained on context windows of Mel Frequency Cepstral Coefficients (MFCC). The best result was achieved using a weighted combination of HMM architectures with 1, 3, and 5 states, where 1-state HMM gave the highest individual performance. In addition to its classification capability, this model is appropriate for our work because it can model local dynamics and aggregate them into global predictions, thus enabling both utterance and sub-utterance data selection.

We make two modifications to this model to make it more suited for our work. Firstly, because the main focuses of this paper are on the analysis and interpretation of classification results, we do not combine different HMM architectures and instead use 1-state HMM exclusively. We find that this model strikes a good balance between classification performance and ease of analysis. Secondly, we use a 1-layer Artificial Neural Network (ANN) for estimating emission probabilities instead of a 5-layer DBN. Because our training set will be constantly refined as data instances are gradually added, the DBN's generative pretraining step will be too computationally expensive. Our preliminary experiments indicate that 1-layer ANNs perform comparably to 5-layer DBNs when using the full training set. It is possible that the FAU Aibo 2-class problem has a simpler decision boundary, thus using a more complex model does not provide any significant advantage.

Algorithm 1: Data selection experiment

Data:
$sl \in \{UCB, SUCB, SUS\}$: selection level
$rm \in \{A, D, M, U\}$: ranking method
$\mathcal{I}$: set of initial utterances to train the first classifier
$\mathcal{D}$: set of training data instances to select from
$\mathcal{N}$: number of data instances to select per iteration
Result:
$\mathcal{T}$: time series of test UARs

```
TrainingSet = I
T = {}
while D is not empty do
    Train new classifier C using TrainingSet
    Compute test UAR using C and add this UAR to T
    Calculate classifier agreement for each instance in D
    Rank classifier agreement according to rm
    Remove top N data instances from D in accordance
    with sl and add them to TrainingSet
end
Train classifier C using the full training set
Compute final test UAR using C and add this UAR to T
return T
```

B. Data Selection Experiment: High-level Overview

Algorithm 1 gives an overview of our data selection experiment. The algorithm selects a small batch of data instances to add to the training set at each iteration. The parameters that we control for are the **selection level** sl and the **ranking method** rm. The former determines the unit on which data selection is performed, while the latter governs how data instances are ranked. Both parameters will be discussed in more detail shortly. Remaining inputs, namely the initial utterance set $\mathcal{I}$, the training data pool $\mathcal{D}$, and the selection size $\mathcal{N}$, are hyperparameters. Section V describes how they are set.

C. Selection Metric: Classifier Agreement

An ideal **selection metric** should convey meaningful characteristics of a data instance and help predict its efficacy when added to the training set. A metric widely used in previous works is classifier agreement, i.e. the degree to which a classifier agrees with the ground-truth. Intuitively, selecting instances with significantly high agreement will have little impact on the decision boundary, leading to small but predictable performance gain. In contrast, choosing instances with low agreement may drastically alter the decision boundary, producing large and erratic change. Since emotion perception and expression are highly varied, instances with very low agreement may represent several underlying classes and induce classifiers not suited to handle the diversity of a given emotion. Excluding such data may help facilitate the training process.

The level of agreement for a given utterance $\mathbf{U}$ can be easily derived from a HMM-ANN classifier. Let $P(\mathbf{U}|IDL)$ and $P(\mathbf{U}|NEG)$ be the probability that the HMM for emotion IDL and NEG generates $\mathbf{U}$, respectively. Let $L_{\mathbf{U}} \in \{IDL, NEG\}$ be the emotion label assigned to $\mathbf{U}$. All ground-truth labels are assumed to be known because we are analyzing the data selection component in isolation. These labels would not be available in an active learning framework. The classifier agreement for $\mathbf{U}$ can then be calculated as:

$$CA(\mathbf{U}) = \frac{P(\mathbf{U}|L_U)}{P(\mathbf{U}|IDL) + P(\mathbf{U}|NEG)} \in [0, 1] \quad (1)$$

Classifier agreement can also be computed at the sub-utterance level. In HMM-ANN modeling, utterance $\mathbf{U}$ is represented as a sequence of context windows (a.k.a frames) $\{W_1, W_2, \ldots, W_n\}$, each of which spans a short segment within the utterance. Let $P(W_i|IDL)$ and $P(W_i|NEG)$ be the probability that the HMM state associated with emotion IDL and NEG emits W_i, respectively. Let L_{W_i} be the emotion label of the utterance containing this context window. The classifier agreement for W_i can then be calculated as:

$$CA(W_i) = \frac{P(W_i|L_{W_i})}{P(W_i|IDL) + P(W_i|NEG)} \in [0, 1] \quad (2)$$

For both utterance and sub-utterance agreement, a score of 1 means total consensus with the ground-truth, while a score of 0 means complete lack of consensus.

Another potential selection metric explored in [13] is the level of agreement between human annotators. This metric has been shown to be effective, but is not straightforward to apply to sub-utterance selection because it is usually assigned at the utterance level. Our preliminary experiments indicate that human and classifier agreement do not produce significantly different results. Due to space constraint, we focus exclusively on classifier agreement in this work.

D. Selection Level: Utterance vs. Sub-Utterance

The traditional way to select data, which was adopted by most related works, is to choose which utterances to include in the training set. The importance of maintaining class balance and making the minority label more salient has been demonstrated in [13]–[15]. In this work, we maintain class balance using a scheme named ***Utterance Class Balancing (UCB)***, which distributes the number of utterances to select among the two emotion classes. For example, suppose the current training set has 400 IDL and 300 NEG utterances, and we want to select 200 additional utterances. 50 IDL and 150 NEG instances will then be chosen, making the training set balanced at 450 utterances per class.

As discussed in [2], it may be the case that only a short segment within an utterance actually influences human perception of certain emotion. In such cases, considering all context windows in an utterance might introduce too much noise and mask the salient local pattern. We therefore propose two methods for sub-utterance data selection to address this problem. In these methods, only a subset of context windows within an utterance is chosen to add to the training set at each iteration. We call the first method ***Sub-Utterance Class Balancing (SUCB)***. It is similar to *UCB*, except that the class labels are counted at the context window level instead. For example, suppose the

current training set has 3,000 IDL and 2,500 NEG context windows, and we want to select 1,500 additional windows. 500 IDL and 1,000 NEG instances will therefore be chosen, making the training set balanced at 3,500 windows per class. A possible drawback of *SUCB* is that some utterances may contribute a lot of context windows, while others may supply none at all. If there is indeed some signal within all utterances, it might be beneficial to ensure that every utterance contributes to the training set. This leads to our second selection method named ***Sub-Utterance Sampler (SUS)***. *SUS* chooses a number of context windows from every utterance to add to the training set, where the selection size is proportional to the length of the utterance. The key difference between *SUCB* and *SUS* is that in the former, context windows from all utterances are pooled together for selection, whereas in the latter, windows are selected separately for each utterance.

E. Ranking Method: Ascending, Descending, Median, Uniform

Once the selection level and metric are specified, the last step is to decide how data instances are ranked for selection. In this work, we consider four strategies for ranking data instances. ***Ascending (A)*** ranks instances in increasing agreement rates, ensuring that those with the lowest consensus with the ground-truths are selected first. In contrast, ***Descending (D)*** ranks instances in decreasing agreement rates, favoring data points with the highest consensus. ***Median (M)*** ranks instances by how close they are to the median position, therefore preferring those in the middle of the agreement spectrum. Finally, ***Uniform (U)*** selects instances at equally-spaced intervals across the entire agreement spectrum. Each of these methods has its own pros and cons and encodes our intuition about data selection. For example, ranking method *D* will likely result in steady but slow performance gain, while method *A* is expected to produce large but erratic changes in performance. Strategies *M* and *U* aim to find the middle ground between these two methods. *M* attempts to strike a balance between stability (high agreement) and performance gain (low agreement). Meanwhile, *U* tries to diversify training data by selecting from all parts of the agreement spectrum. Compared to *M*, ranking method *U* may result in higher performance gain but lower stability. Analyzing the impact of each ranking method will provide valuable insights into the type of data that is most beneficial for the training process.

V. Experiments

Features: We adopt a similar feature set used in [6]. We extract 12-dimensional MFCCs plus energy along with delta and delta-delta coefficients for each utterance, using a 25ms Hamming window and 10ms time shift. We then create a sequence of 10-frame context windows centered around each frame, padding out-of-boundary frames with zeros if necessary. Training features are z-normalized at the speaker level. Test features are globally z-normalized using the test set's mean and standard deviation. Previous experiments have shown that low-level acoustic features like MFCC are more suitable for dynamic frame-level emotion modeling compared to utterance-level statistics commonly used in static classification, such as the INTERSPEECH ComParE feature set [22].

ANN Training: We adopt the network architecture and finetuning process used in [6]. We use a 1-layer ANN with 1024 hidden units. Six random training speakers are withheld to form the validation set. The IDs of these speakers are: *Ohm_06*, *Ohm_07*, *Ohm_11*, *Ohm_14*, *Ohm_16*, and *Ohm_20*. The resulting training and validation sets contain 8,269 and 1,690 utterances, respectively. We train the network using backpropagation with $\tanh$ activation, 0.9 momentum, 2×10^{-5} L2 regularization coefficient, and an initial learning rate of 0.01. After each training epoch, if the validation UAR decreases, we restore the previous network weights, halve the learning rate, and resume the training process. This is repeated until the learning rate falls below 0.0001. Training is significantly sped up using Theano with GPU support [23].

Data Selection: Our data selection experiments adopt the parameters used in [14], [15], [17]. The initial training set $\mathcal{I}$ contains 500 randomly chosen utterances. The selection size $\mathcal{N}$ is set at 200 utterances and 36,000 context windows for utterance and sub-utterance selection, respectively. With this setup each experiment terminates after 40 epochs, where the last iteration involves all training data. For each parameter combination of selection level and ranking method, we repeat the experiment 10 times with different random seeds to help analyze the method's stability and determine the point of convergence. The randomness comes from the shuffling of training data after each backpropagation epoch to avoid overfitting.

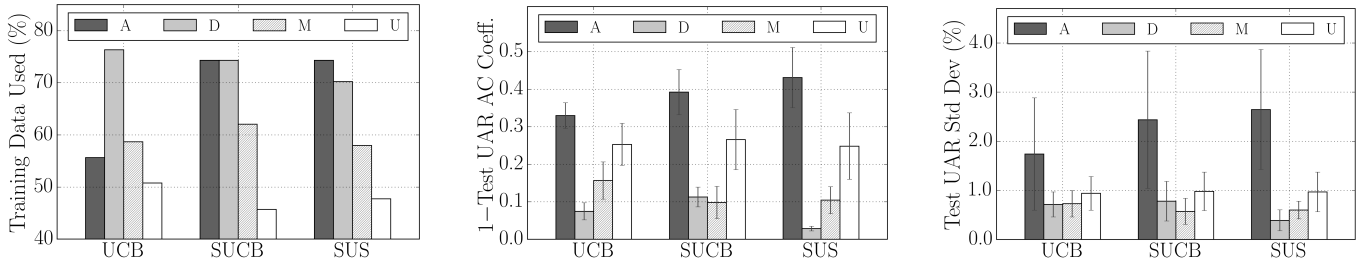
VI. Results and Discussion

A. Classification Performance

The average test UAR of our hybrid HMM-ANN classifier using the full training set, computed over 120 independent runs (3 selection levels, 4 ranking methods, 10 repetitions), is $\mathbf{70.47\% \pm 0.58\%}$, with a median UAR of $\mathbf{70.68\%}$. These results are comparable to that achieved by the winner of the 2009 emotion challenge (70.3% UAR), suggesting that our simple baseline model has good classification capability. In this section we analyze the evolution of test set UAR as data instances are incrementally selected for training. Specifically, we are interested in two aspects: (1) the amount of data required to achieve the same performance as the full training set and (2) the stability of the learning process.

1) Convergence Rate: We say an experiment has converged if the test UARs achieved by the classifier at a given iteration (sample size: 10) are not statistically significantly different from those achieved when using the full training set (sample size: 120). Because the final test UARs are not well approximated by the normal distribution and are skewed to the left, we use Mood's median test ($p = 0.05$) to evaluate if the medians of these two samples are identical. This nonparametric test makes no assumption about a sample's distribution and is robust against outliers, making it suitable in this scenario.

Figure 1(a) shows the amount of training data used by each selection method at the point of convergence. This is defined as the fraction of context windows currently in the training



(a) Point of convergence (b) Variation within single runs (c) Variation across multiple runs

Fig. 1: Properties of test set UARs. Comparing different **selection levels**: Utterance Class Balancing (*UCB*), Sub-Utterance Class Balancing (*SUCB*), Sub-Utterance Sampler (*SUS*) and **ranking methods**: Ascending (*A*), Descending (*D*), Median (*M*), Uniform (*U*). Point of convergence (a) is determined by the Mood's median test ($p = 0.05$).

set. Since utterances are varied in length, we report the mean and standard deviation of the amount of context windows used for *UCB*. To make the results more comparable with previous works, we calculate the percentage of data over the full training set plus the held-out validation set. The best convergence rate is achieved with the *Uniform* ranking method in all cases. For the *UCB*, *SUCB*, and *SUS* selection level, the amount of training data used at convergence is $50.8\% \pm 0.22\%$, 45.7%, and 47.8%, respectively. With the exception of *UCB*, the *Median* ranking method produces the second best convergence rate, followed by *Descending* and *Ascending*, respectively. As expected, the *Descending* ranking method leads to a slow learning rate, a result of its "safe" selection strategy of choosing data instances with the highest agreement. *Ascending* produces convergence points significantly worse than *Median* and similar to *Descending* for both sub-utterance selection methods. Interestingly, the convergence point for *Ascending* in utterance selection is similar to that achieved by *Median*. We hypothesize that because even utterances with low agreement may contain salient context windows, training the classifier on them still results in reasonable convergence.

Sub-utterance selection, particularly *SUS*, has slightly better convergence than utterance selection for *Descending*, *Median*, and *Uniform*. We hypothesize that the difference in convergence rate is not extremely high due to FAU Aibo's manual segmentation. The theoretical advantage of sub-utterance selection is that it can identify emotionally salient regions within an arbitrarily long utterance. Because utterances in FAU Aibo are pre-segmented chunks with relatively homogeneous emotion, sub-utterance selection does not lead to significantly faster learning on this dataset. In future work we will investigate if applying sub-utterance selection to datasets with longer and less finely-segmented units would lead to larger gain.

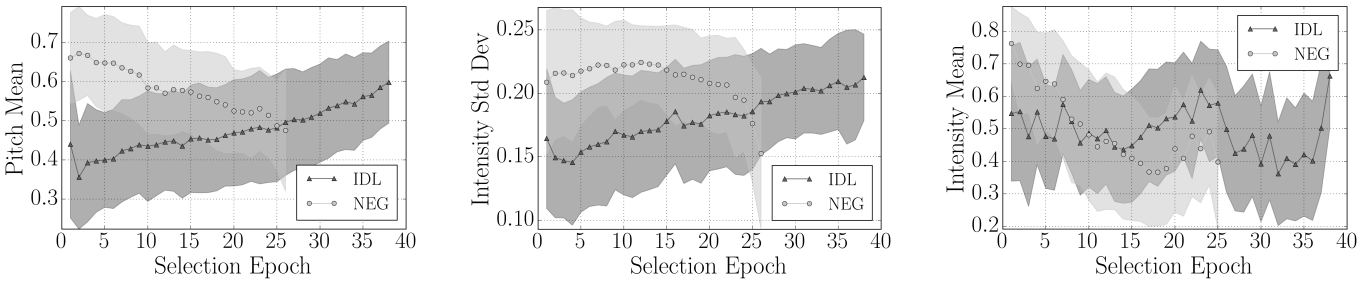
2) Stability: Convergence rate is not the only metric of interest when evaluating a data selection strategy. Also of great importance is the stability of the training process. For certain applications, especially those involving a large amount of data from heterogeneous sources, it is sometimes preferable to sacrifice the rate of learning for a more stable and predictable system, as opposed to a classifier that varies widely in performance at different iterations. In such cases, it is crucial to establish a reasonable trade-off between the rate and stability of learning. In this section we analyze two types of stability: within individual experiments and across multiple runs.

The stability within individual experiments is characterized by the smoothness of the UAR time series. We capture this property by computing the series' lag-one autocorrelation coefficient. The result is a number between -1 and 1. A value close to 1 implies a smoothly varying series with strong linear relationship between two adjacent points. A value close to 0 implies no overall linear relation. Finally, a value close to -1 implies a series heavily jagged around the mean. Figure 1(b) shows the amount of variation present in single runs for different selection methods, measured by subtracting the resulting coefficient from 1. The overall trend is that *Ascending* has the highest within-run variation, followed by *Uniform*, *Median*, and *Descending*, respectively. The figure demonstrates a drawback of using *Uniform* as the ranking method: its fast convergence rate comes with a cost of lower stability. On the other hand, *Median* is more stable, but converges more slowly. The *Ascending* ranking method in *UCB* has less within-run variation than in *SUCB* and *SUS*. As hypothesized previously, this is possibly due to the fact that even utterances with low agreement can contain emotionally salient context windows. Lastly, sub-utterance selection, specifically *SUS*, has significantly lower within-run variation than utterance selection for the *Descending* and *Median* ranking methods.

We also analyze the amount of variation across multiple runs, measured by the standard deviation of test UARs at the same time steps. As shown in Figure 1(c), the trend observed in within-run variation can also be seen here. Specifically, *Ascending* is the most unstable ranking method, more so in sub-utterance selection. *Uniform* shows more variation than *Median*. Finally, *SUS* demonstrates higher stability than *UCB* for both the *Median* and *Descending* ranking methods.

B. Characteristics of Selected Data

In this section, we explore the properties of selected data instances to provide insights into the emotion classifier as well as the data selection algorithm. We restrict our analysis to setups that use the *Descending* ranking method, allowing us to study changes as a function of decreasing classifier agreement.



(a) *UCB-Descending* pitch mean (b) *UCB-Descending* intensity variation (c) *SUCB-Descending* intensity mean

Fig. 2: Acoustic properties of utterances selected by *UCB-Descending* and context windows selected by *SUCB-Descending*. In all graphs, classifier agreement decreases along the x-axis. Here we are selecting the same number of samples for IDL and NEG at each iteration to maintain class balance. After 25 epochs there is no more data to select for NEG, the minority class.

1) Acoustic Properties: Two acoustic properties commonly tied to emotion recognition from speech are pitch and intensity [24]. In order to match the normalization method used on training data, we z-normalize the pitch and intensity of the training utterances at the speaker level, followed by sigmoid normalization to constrain the values between 0 and 1. A pitch/intensity value of over 0.5 means higher than average and vice versa. Finally, we compute the mean and standard deviation of pitch and intensity for each selected utterance and context window.

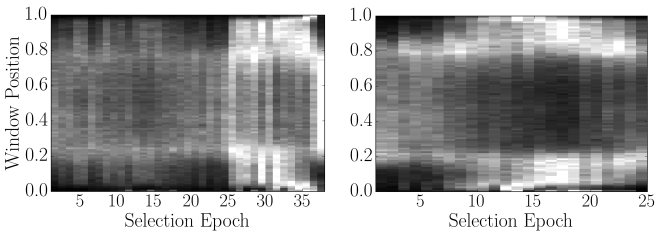
As shown in Figure 2(a), as classifier agreement decreases, the average utterance-level pitch value for *IDL* and *NEG* steadily increases and decreases, respectively. This suggests that pitch mean is a good indicator of emotion prototypicality, and data instances with "abnormal" pitch value, i.e. high for *IDL* and low for *NEG*, are harder to classify. A similar trend (not shown) can be observed in *SUCB* at the window level.

Figure 2(b) shows that the utterance-level standard deviation of intensity for *IDL* is negatively correlated with classifier agreement, while the trend for *NEG* is not as clear. Intuitively, this means that emotion *IDL* can typically be characterized as having monotonic loudness. This description matches *IDL*'s role in FAU Aibo as a quasi-neutral class.

Finally, the mean intensity of selected context windows for emotion *NEG* is positively correlated with classifier agreement, as seen in Figure 2(c). Interestingly, utterance selection does not exhibit such a trend. This suggests that the prototypical *NEG* utterances have short bursts of intensity which can be captured more effectively by sub-utterance selection.

The results shown here, combined with the fact that the *Descending* ranking method leads to slow convergence, imply that it is better to select instances with diverse acoustic profiles.

2) Relative Window Position: To better understand sub-utterance selection, we plot the relative position within an utterance of selected context windows over time. Figure 3 shows that for emotion *NEG*, the classifier selects primarily windows close to the center first and leaves those near the boundaries for later iterations. This provides a general idea of the location in which emotion is expressed. The trend is not as clear for *IDL*, suggesting that this emotion is expressed in a less localized manner. These findings echo the results in [25], which showed that good classification performance can be achieved using the middle third of an utterance. Unlike in their work, we did not pre-segment the utterances into thirds.



(a) IDL (b) NEG

Fig. 3: Relative position of context windows selected by *SUCB-Descending*. Lighter regions mean higher occurrences.

VII. Conclusion and Future Work

In this work, we provide a comprehensive review and analysis of various utterance and sub-utterance data selection methods for acoustic emotion recognition in the context of a hybrid HMM-ANN classifier. We demonstrate that the *Uniform* ranking method provides the best convergence rate, achieving a median UAR of 70.68% on FAU Aibo's 2-class problem with less than 50% of the training data. On the other hand, *Median* has slower convergence but higher stability. Sub-utterance selection results in slightly faster learning rate compared to utterance selection and is more stable both within single and across multiple runs. Finally, we show that the selected data instances have intuitive interpretations based on their pitch and intensity profiles, as well as the segments' relative positions within the utterance.

For future work, we plan to expand the analysis to a multi-corpus setting to see if our methods can leverage large amount of heterogeneous emotional speech data. We will explore if sub-utterance selection provides more advantage on datasets with coarser units. We will also tackle semi-supervised training by selecting data from unlabeled sources. Lastly, we will experiment with a system that combines classifier and human agreement, and adaptively selects the optimal ranking method.

REFERENCES

[1] S. Steidl, "Automatic Classification of Emotion Related User States in Spontaneous Children's Speech," Ph.D. dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2009.

[2] C.-C. Lee, A. Katsamanis, P. Georgiou, and S. S. Narayanan, "Based on Isolated Saliency or Causal Integration? Toward a Better Understanding of Human Annotation Process using Multiple Instance Learning and Sequential Probability Ratio Test," in *Proc. of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, 2012.

[3] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing," in *Proc. of the 2nd International Conference on Affective Computing and Intelligent Interaction (ACII)*. Lisbon, Portugal: Springer-Verlag, 2007, pp. 139–147.

[4] E. Mower and S. S. Narayanan, "A Hierarchical Static-Dynamic Framework For Emotion Classification," in *Proc. of the IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011.

[5] E. Mower Provost, "Identifying Salient Sub-Utterance Emotion Dynamics Using Flexible Units and Estimates of Affective Flow," in *Proc. of the IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 2013.

[6] D. Le and E. Mower Provost, "Emotion Recognition from Spontaneous Speech using Hidden Markov Models with Deep Belief Networks," in *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Olomouc, Czech Republic, 2013, pp. 216–221.

[7] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brighton, United Kingdom, 2009, pp. 312–315.

[8] C. E. Erdem, E. Bozkurt, E. Erzin, and A. T. Erdem, "RANSAC-Based Training Data Selection for Emotion Recognition from Spontaneous Speech," in *Proc. of the 3rd International Workshop on Affective Interaction in Natural Environments (AFFINE)*. New York, NY, USA: ACM, 2010, pp. 9–14.

[9] E. Bozkurt, E. Erzin, C. E. Erdem, and A. Erdem, "RANSAC-Based Training Data Selection on Spectral Features for Emotion Recognition from Spontaneous Speech," in *Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issues*. Springer Berlin Heidelberg, 2011, vol. 6800, pp. 36–47.

[10] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[11] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Selecting Training Data for Cross-Corpus Speech Emotion Recognition: Prototypicality vs. Generalization," in *Proc. of the Afeka-AVIOS Speech Processing Conference*, Tel Aviv, Israel, 2011.

[12] S. Meudt and F. Schwenker, "On Instance Selection in Audio Based Emotion Recognition," in *Artificial Neural Networks in Pattern Recognition*. Springer Berlin Heidelberg, 2012, vol. 7477, pp. 186–192.

[13] Z. Zhang, F. Eyben, J. Deng, and B. Schuller, "An Agreement and Sparseness-based Learning Instance Selection and its Application to Subjective Speech Phenomena," in *Proc. of the 5th International Workshop on Emotion Social Signals, Sentiment & Linked Open Data (ES^3 LOD)*, Reykjavik, Iceland, 2014.

[14] Z. Zhang and B. Schuller, "Active Learning by Sparse Instance Tracking and Classifier Confidence in Acoustic Emotion Recognition," in *Proc. of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, 2012, pp. 362–365.

[15] Z. Zhang, J. Deng, E. Marchi, and B. Schuller, "Active Learning by Label Uncertainty for Acoustic Emotion Recognition," in *Proc. of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Lyon, France, 2013, pp. 2856–2860.

[16] S. Mariooryad, R. Lotfian, and C. Busso, "Building A Naturalistic Emotional Speech Corpus by Retrieving Expressive Behaviors From Existing Speech Corpora," in *Proc. of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore, 2014, pp. 238–242.

[17] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative Learning and Its Application to Emotion Recognition from Speech," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 23, no. 1, pp. 115–126, Jan. 2015.

[18] P. Dumouchel, N. Dehak, Y. Attabi, R. Dehak, and N. Boufaden, "Cepstral and Long-Term Features for Emotion Recognition," in *Proc. of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brighton, United Kingdom, 2009.

[19] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge," *Speech Communication*, vol. 53, no. 910, pp. 1062 – 1087, 2011.

[20] A. Hassan, R. Damper, and M. Niranjan, "On Acoustic Emotion Recognition: Compensating for Covariate Shift," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 7, pp. 1458–1468, 2013.

[21] Y. Bengio, "Learning Deep Architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, Jan. 2009.

[22] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. R. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism." in *Proc. of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Lyon, France, 2013, pp. 148–152.

[23] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU Math Expression Compiler," in *Proc. of the Python for Scientific Computing Conference (SciPy)*, Austin, TX, USA, 2010.

[24] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of Emotion Recognition Using Facial Expressions, Speech and Multimodal Information," in *Proc. of the 6th International Conference on Multimodal Interfaces (ICMI)*. New York, NY, USA: ACM, 2004, pp. 205–211.

[25] B. Schuller and G. Rigoll, "Timing Levels in Segment-Based Speech Emotion Recognition." in *Proc. of the 9th International Conference on Spoken Language Processing (INTERSPEECH)*, Pittsburgh, PA, USA, 2006.