Domain Adaptation and Compensation for Emotion Detection

Michelle Hewlett Sanchez[1,2], *Gokhan Tur*[3]*, *Luciana Ferrer*[1], *Dilek Hakkani-Tür*[4]

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA 94025, U.S.A.
[2]Stanford University, Stanford, CA 94305, U.S.A.
[3]Speech at Microsoft | Microsoft Research, Mountain View, CA 94041, U.S.A.
[4]International Computer Science Institute (ICSI), Berkeley, CA 94704, U.S.A.
{mhewlett,lferrer}@speech.sri.com, {gokhan.tur,dilek}@ieee.org

Abstract

Inspired by the recent improvements in domain adaptation and session variability compensation techniques used for speech and speaker processing, we study their effect for emotion prediction. More specifically, we investigated the use of publicly available out-of-domain data with emotion annotations for improving the performance of the in-domain model trained using 911 emergency-hotline calls. Following the emotion detection literature, we use prosodic (pitch, energy, and speaking rate) features as the inputs to a discriminative classifier. We performed segment-level n-fold cross validation emotion prediction experiments. Our results indicate significant improvement of performance for emotion prediction exploiting out-of-domain data.

Index Terms: emotion detection, domain adaptation

1. Introduction

In the framework of a larger-scale project on speech-based situational awareness for emergency response, we are interested in predicting the emotional status of 911 emergency-hotline callers. There is an established body of research attempting to characterize the emotional state of human speech, such as anger, fear, joy, and sadness. However the largest part of the existing literature relies on acted emotional speech with high signal-to-noise ratios, as in the Linguistic Data Consortium (LDC) Emotional Prosody Speech and Transcripts Corpus.

This paper discusses our methods and results for the emotion detection of fear versus neutral in a 911 emergency call setting. The main focus of this paper is exploring novel domain adaptation and compensation methods to exploit the out-of-domain data, in our case the LDC Corpus. The first approach is based on established domain adaptation techniques. Although statistical model adaptation has been a well studied area in speech and language processing (such as language modeling [1], call classification [2], and dialog act tagging [3]) there is no previous study on emotion model adaptation.

The second approach is based on recent breakthrough improvements in speaker processing via session variability compensation, more specifically nuisance attribute projection (NAP), where typically the nuisance is the acoustic conditions, such as recording device [4, 5] or emotion [6]. We use the NAP techniques to propose a novel way of viewing the session variability compensation in which the *domain* is treated as a kind of *nuisance* which needs to be compensated for. To the best of our knowledge, there are no speech and language processing studies attempting to build models robust for different domains to predict emotion via this technique.

In the next section, we review the emotion detection literature briefly. Section 3 briefly describes the modeling approach for emotion prediction using prosodic features. Then in Section 4, we present the domain adaptation and compensation methods studied. Section 5 describes the experimental setup using the 911 calls manually transcribed and annotated and presents the results. Section 6 discusses future work.

2. Emotion Detection

Much research on emotion detection has been done on changes that are seen in acoustic features, such as speaking rate, pitch, and intensity. Liscombe *et al*. extracted acoustic features consisting of pitch and energy features and also added some hand-labeled features [7]. They used a binary classification algorithm to differentiate between 10 different emotions by detecting the presence or absence of each emotion. Bhatti *et al*. classified six emotions (happiness, sadness, anger, fear, surprise, disgust) using a neural network approach [8].

Prosodic features such as pitch and energy, mean pause length, and speaking rate were extracted before sequential forward selection was performed to select the most useful features. Lugger and Yang used the Berlin Emotional Speech Database to classify six emotions (anger, happiness, sadness, boredom, anxiety, neutral) with a Bayesian classifier modeled with Gaussians using both prosodic and voice quality features [9]. Schuller *et al*. classified six different emotions (same as in [9]) based on a combination of a Gaussian mixture model (GMM) and a hidden Markov model (HMM) approach using both acted and real emotional speech in both English and German, and using pitch and energy features [10]. Schuller *et al*. have done recent work on the influence of noise and microphone conditions on both acted and spontaneous data [11]. The acted speech consisted of two databases, the Danish Emotional Speech Corpus and Berlin Emotional Speech Database, and the spontaneous speech was from a German corpus of recordings of children communicating with a robot controlled by a human. They used decision trees to classify emotion based on two different feature sets including pitch, energy, and spectral features, mel frequency scaled-cepstral coefficients (MFCCs), and additional features.

Past research performed on non-acted speech data included the following: Ang *et al*. employed human-computer dialog strictly used for research making feigned air travel reservations in order to classify annoyed and frustrated versus neutral using decision trees [12]. Their prosodic model included features such as pitch, energy, pause, duration, speaking rate, and spec-

*This work was done while the author was at SRI International.

tral tilt. Liscombe *et al*. used data from the AT&T "How May I Help You" spoken dialog system to detect negative versus nonnegative emotions using a type of boosting algorithm [13]. Their feature space included both lexical features using a trigram model and prosodic features including pitch, energy, voice quality, and speaking rate. Lee *et al*. extracted acoustic features like pitch and energy and performed principal component analysis to reduce the feature space while maximizing the recognition accuracy [14]. Lee and Narayanan added lexical and discourse features into both a linear discriminant classifier and a k-nearest neighborhood classifier [15] . Devillers and Vidrascu have done work similar to our task using a real-life data corpus of calls from a medical emergency call center [16]. Their task was to classify relief, anger, fear, and sadness using both linguistic and paralinguistic features. The linguistic features were based on a unigram model, and the paralinguistic features included pitch, spectral, energy, and duration.

3. Modeling Emotion

As established in the literature [12, 16, among others], we exploited prosodic features for emotion detection. As the emergency call center data we are working with had low audio quality and the spoken utterances were very emotional and ungrammatical, in order not to deal with the speech recognition errors, the experiments have been carried out using the manual transcriptions. The prosodic feature are extracted using the SRI Algemy toolkit from the forced alignment of the speech to manual transcriptions for detecting the start times of individual words. Algemy contains a graphical user interface that allows users to easily read and program scripts for the calculation of prosodic features using modular algorithms as building blocks. These blocks are strung together in directed acyclic graphs to extract the desired features.

Motivated by previous work, we extracted the following types of prosodic features for each segment: pitch (extracted as the fundamental frequency (F0)), energy (extracted as the root mean square (RMS) of the amplitude of the signal), and speaking rate. Pitch features included mean, maximum, minimum, range, and standard deviation of the voiced segments of logarithm of F0. In most previous work, log F0 is normalized per speaker. Each speaker appeared in only one phone call so this was not an option for us. Instead of being normalized over the entire data set, the pitch features were post-processed and normalized globally based on gender.

Energy features included mean, maximum, minimum, range, and standard deviation of log energy normalized per call for the range of 0 to 1. The speaking rate was the number of phones divided by the duration of the segment. The speaking rate was normalized globally over all the calls

The features, as extracted from the segments, are then fed into a support vector machine (SVM) classifier. While the modeling is always at the segment level, as the manual emotion annotation has been done at the call level, the segments are assigned to the emotion label of the calls during training.

The automatically annotated segments are then postprocessed. The majority emotion category in each call is assigned as the predicted emotion of that call. The evaluation is then done at the segment level using the F-Measure metric on the minority class, i.e., fear emotion.

4. Domain Adaptation and Compensation

In our work, we explored the use of out-of-domain emotion data, namely the LDC Corpus, when testing the 911 data in two different ways: domain adaptation and domain compensation.

4.1. Domain Adaptation

In cases where only a limited amount of data annotated with emotion is available, an immediate solution would be to use supervised domain adaptation methods with existing out-of-domain data or models.

Based on the taxonomy of Dasarathy [17], the model adaptation methods can be categorized depending on where the combination happens. Some popular methods include the *feature-in-decision-out* (early) fusion, where the feature space or data is concatenated, *decision-in-decision-out* (late) fusion, as typically done using interpolation, or *decision-as-feature*, where the decision of the out-of-domain model is used as a feature for the in-domain model. We briefly describe these three methods below.

The simplest way of early fusion is data concatenation, where the data from the existing corpus is used as additional training data along with the current corpus. In this study, the LDC Corpus is used as it already had segments labeled with different types of emotions. We used the panic and neutral segments from this corpus as added training data.

In the late fusion, *decision-in-decision-out* ($C^o \times C^i$), a final model combining the decisions of the out-of-domain, C^o, and in-domain, C^i, models is interpolated using some held-out set. In this study we did not employ this method as we did not have a separate held-out set to train the combination model.

In the third method, *decision-as-feature* ($C^o \rightarrow C^i$), the classifier trained using only out-of-domain data, C^o, is run first on the in-domain data. The probability it outputs for each class is then used as an extra feature while training a model with the in-domain data, C^i. The final decision is made by C^i trained on this enriched set of features. We have experimented with these methods in our earlier work on dialog act segmentation [18]. To experiment with the decision-as-feature method, we trained the out-of-domain model using the LDC emotion corpus and used the prediction values from this classifier as an added feature in a new model classifier.

4.2. Domain Compensation

Nuisance compensation for the most part has been used for speaker recognition where the method is used to account for session variability effects due to speech recorded on different channels, with different styles, or with different background noise. Using the Nuisance Attribute Projection (NAP) methods described by Solomonoff *et al.* [4, 5] for more accurate speaker recognition, we account for these nuisance differences to produce more accurate emotion recognition. The nuisance we were compensating for was the domain of the data (911 or LDC) including the channel (911 was telephone speech and LDC was microphone speech) and the length of the speech segments (LDC segments were much shorter than 911). The emotion was either fear or neutral. Throughout the rest of the paper, we call our technique domain compensation since the nuisance is the difference in the domains.

In order to reduce the domain effects using NAP, one needs to create a matrix P to project points in the original space to a space that is more robust to effects from the undesired variability due to domain differences. We used Solomonoff's method

caller: help please
dispatcher: fire emergency
caller: okay my girlfriend just fell over she's like having a seizure i'm so scared somethings going wrong
dispatcher: hello
caller: my girlfriend
dispatcher: hello
caller: yes
dispatcher: what is the address
caller: *address** street apartment l seven oh no i'm sorry not apartment l seven it's the it's the office oh my god her eyes are rolling back
dispatcher: okay calm down okay
caller: okay
dispatcher: what city are you in
caller: i hope she's breathing
dispatcher: what city
caller: santa ana heights
dispatcher: okay, the ambulance is on its way

Figure 1: Sample emotional dialog from 911 corpus. **address* is italicized to keep this information private.

using the kernel space to reduce the domain effects between the 911 and LDC data. First, we create a matrix A whose columns are vectors from the training corpus. Then, X is defined to be

$$X = AV \tag{1}$$

where V is a matrix containing the eigenvectors corresponding to the largest eigenvalues of the symmetric eigenvalue problem

$$KZKV = KV\Lambda \tag{2}$$

where $K = A^T A$. $Z = \texttt{diag}(W\mathbf{1}) - W$ is a matrix that depends on a weight matrix W where

$$W = \alpha W_{domain} - \gamma W_{emotion} \tag{3}$$

and **1** is the vector of all ones. $\texttt{diag}(Y)$ is a matrix whose diagonal elements are the values in the vector Y. The value of an entry in W_{domain} is 1 if the domains are different and the value of an entry in $W_{emotion}$ is 1 if if the emotions are different. In our experiments, we made both α and γ 1 which makes the weight matrix 1 for training points we want to pull together, -1 for training points we want to pull apart, and 0 for the pairs that do not matter. After solving for V, we can solve for X and find the projection matrix $P = I - XX^T$. We then project the original data on this matrix P to get the new domain compensated data. In doing this, we reduce the domain effects and create a bigger distance between the fear and neutral samples.

5. Experiments and results

We performed experiments at the segment level using call-level emotion categories. We explored the use of out-of-domain emotion data, namely the LDC corpus, for this task for both cases. SVM Light with default parameters was used in all our classification experiments in this section [19]. Because we had only 95 phone calls, we used 95-fold leave-one-out-cross-validation to maximize the use of our data. We focused only on fearful and neutral speech, ignoring the sad emotion cases.

	Number of Segments		Number of Calls	
Domain	LDC	911	LDC	911
Male	96	585	3	42
Female	165	657	5	50
Neutral	114	839	-	67
Fear/Panic	147	403	-	28

Table 1: The characteristics of the LDC and 911 data used in the experiments.

5.1. Speech data and emotion labeling

All 911 data was recorded mono over the telephone. The dispatcher's duty is to determine the type of emergency, the caller's location and telephone number, and give advice on what to do before the paramedics arrive. Figure 1 shows the beginning of a transcription from an emotional example in the corpus.

Every phone call was labeled by five different labelers with one of three emotions of the caller: fear, sadness, and neutral. A majority vote was taken to verify emotion agreement. The kappa is 65% between three annotators.

Of the 99 phone calls in the 911 data set, 28 were classified as fear, 4 as sadness, and 67 as neutral. Since so few calls were classified as sadness, we decided to focus on classifying fear versus neutral, giving us a total of 95 phone calls for our experiments. Each phone call was broken down into segments of the caller's voice. These segments were obtained by a turn in the conversation or a natural break in the speech of the caller. Each call had about 13 segments on average. Eight calls had two different caller voices because the original caller gave the phone to a different person during the call and 3 calls containing both male and female voices. Table 1 provides more details about the LDC and 911 data used in the experiments. LDC corpus is more balanced for neutral and emotional speech while most 911 calls are annotated as neutral (either because a health care professional or some third person is making the call or the callers try to remain calm to better describe the situation).

5.2. Classification Results

While the emotion annotations were at the call level, we performed experiments for each segment of the phone call. The results from these experiments are in Table 2 in term of F-measure of detecting fearful calls. The baseline (B) represents the results when assigning fear or the minority class to all utterances (i.e., 100% recall but low precision). Using just five pitch features (P) normalized by gender, we see significant[1] improvement.

Training the classifier using also the LDC segments (911+LDC) always shows a similar result or an improvement over not using these segments in training. The initial experiments were run by adding the panic and neutral segments from the LDC corpus as additional training segments along with the 911 data. An even larger improvement (5% relative) occurs when we train using only the LDC segments first and then use the prediction as an added feature in a new SVM classifier (LDC $\rightarrow$ 911). The LDC model was trained using the five pitch features only.

As seen in Table 2, training the classifier using domain compensated features of the 911 data (NAP 911) and training the classifier using the domain compensated features of both 911 and LDC data (NAP 911+LDC) show improvement over the original experiments without doing feature projection with

[1] According to the Z-test with 0.95 confidence.

Experiment	911	911+LDC	LDC → 911	NAP 911	NAP 911+LDC
B	49.0 %	49.0 %	49.0 %	49.0 %	49.0 %
P	**64.1 %**	64.0 %	**64.6 %**	**64.5 %**	**64.8 %**
P+SR	63.5 %	**64.2 %**	64.4 %	64.4 %	64.8 %
P+E+SR	63.6 %	63.3 %	64.3 %	64.3 %	64.4 %
Postprocess	**70.9 %**	**73.1 %**	**74.5 %**	**71.1 %**	**72.4 %**

Table 2: *Summary of experimental results for classifying fear versus neutral with and without additional LDC training segments, using LDC model score (LDC → 911), and with and without additional LDC training segments after performing NAP. F-measure is the score used in the table. B is baseline, P is pitch features, SR is speaking rate, E is energy features.*

NAP.

In postprocessing, we labeled all of the segments of each call with the majority decision label and this resulted in significant improvements in the classification performance. We used the best F-measure obtained in each column of Table 2 to calculate the postprocessing results.

Overall, using NAP techniques when training on both 911 and LDC performed the best before postprocessing and using the LDC model score as an additional feature when classifying the 911 calls showed the best results after postprocessing.

6. Conclusions

We have shown that using out-of-domain data improves our results in the 911 emergency call setting. We experimented with both established domain adaptation methods and proposed using domain compensation methods for this purpose, viewing the domain mismatch as a nuisance.

Detecting emotion of real-life data is a very important yet difficult task. Most data sets used in the past have had multiple acted conversations or calls from the same speaker. In our case, we had only one call for each speaker which is harder to model.

In the future, we would like to add more prosodic features like the first derivative of pitch or pause duration to our classification algorithm, or contextual features like whether or not the caller and dispatcher interrupt each other during the call, and how much speech overlap there is in each segment of the call.

7. Acknowledgments

We thank Elizabeth Shriberg and Professor Robert M. Gray for their valuable discussions. We also acknowledge Harry Bratt, and Martin Graciarena who developed the software to extract the features and for their added help in using this software. This research was funded by NSF Grant #IIS-0812610, #IIS-0710833, and #CCF-0846199.

8. References

[1] J. R. Bellegarda, "Statistical language model adaptation: Review and perspectives," *Speech Communication Special Issue on Adaptation Methods for Speech Recognition*, vol. 42, pp. 93–108, 2004.

[2] G. Tur, "Model adaptation for spoken language understanding," in *Proceedings of the ICASSP*, Philadelphia, PA, May 2005.

[3] G. Tur, U. Guz, and D. Hakkani-Tur, "Model adaptation for dialog act tagging," Palm Beach, Aruba, 2006, IEEE Workshop on Spoken Language Technology, pp. 94–97.

[4] A. Solomonoff, C. Quillen, and W.M. Campbell, "Channel Compensation for SVM Speaker Recognition," in *Proceedings of Odyssey*, pp. 57-62, 2004.

[5] A. Solomonoff, W.M. Campbell, and I. Boardman, "Advances in Channel Compensation for SVM Speaker Recognition," in *Proceedings of the ICASSP*, Philadelphia, 2005.

[6] H. Bao, M. Xu, and T.F. Zheng, "Emotion Attribute Projection for Speaker Recognition on Emotional Speech," in *Proceedings of Interspeech*, Antwerp, 2007.

[7] J. Liscombe, J. Venditti, and J. Hirschberg, "Classifying subject ratings of emotional speech using acoustic features," Geneva, Switzerland, 2003, Interspeech.

[8] M.W. Bhatti, Y. Wang, and L. Guan, "A neural network approach for human emotion recognition in speech," Vancouver, Canada, 2004, ISCAS.

[9] M. Lugger and B. Yang, "The relevance of voice quality features in speaker independent emotion recognition," Honolulu, HI, 2007, ICASSP.

[10] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," Hong Kong, 2003, ICASSP, pp. 401–404.

[11] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," Honolulu, HI, 2007, ICASSP.

[12] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," Denver, CO, 2002, ICSLP.

[13] J. Liscombe, G. Riccardi, and D. Hakkani-Tur, "Using context to improve emotion detection in spoken dialog systems," Lisbon, Portugal, 2005, Interspeech.

[14] C.M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," Madonna di Campiglio, Italy, 2001, IEEE Automatic Speech Recognition and Understanding Workshop.

[15] C.M. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[16] L. Devillers and L. Vidrascu, "Real-life emotion recognition in speech," *Speaker Classification II, LNAI 4441*, pp. 34–42, 2007.

[17] B. V. Dasarathy, "Sensor fusion potential exploitation - innovative architectures and illustrative applications," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 24–38, 1997.

[18] S. Cuendet, D. Hakkani-Tür, and G. Tur, "Model adaptation for sentence segmentation from speech," in *Proceedings of the IEEE/ACL SLT Workshop*, Aruba, 2006.

[19] "Svmlight support vector machine toolkit," *http://svmlight.joachims.org*.