Cross-genre training for automatic prosody classification

Anna Margolis[1,2]*, Mari Ostendorf*[1]*, Karen Livescu*[2]

[1]Department of Electrical Engineering, University of Washington, Seattle, WA
[2]TTI-Chicago, Chicago, IL

{amargoli,mo}@ee.washington.edu, klivescu@uchicago.edu

Abstract

We consider methods for training a prosodic classifier using labeled training data from a different genre than the one on which the system will be deployed. Two binary tasks are considered: word-level pitch accent and phrase boundary detection. Using radio news and conversational telephone speech, we consider cross-genre training using acoustic and textual features, and find that acoustic features transfer better than text features in most cases. We also find that a single classifier trained from both genres nearly matches genre-dependent performance. We then consider some simple unsupervised domain adaptation approaches, including class proportion adjustment, sample selection bias correction, and feature normalization. With the exception of class proportion adjustment, which is slightly helpful in one case but proves unstable, none of the approaches improve cross-genre performance over the baseline.

Index Terms: prosody recognition, domain adaptation

1. Introduction

Significant research has been done on automatic detection of symbolic prosody classes in speech, such as [1, 2, 3]. Most of the proposed methods rely on supervised learning, which requires a training set that has been hand-labeled for the classes of interest; the labeling process is time-consuming and expensive, requiring the efforts of a trained linguist. As a result, most researchers make use of a few available labeled corpora, such as the Boston University Radio News Corpus [4]. It is known that prosodic characteristics can vary by genre and style, for instance, between professionally read news speech and spontaneous conversational speech [5, 6, 7, 8]. Conversational speech is generally faster, with more backchannels and disfluencies. It has been shown that read speech has a higher proportion of words with pitch accents [5], and that there are different acoustic and lexical indicators of sentence boundaries between the genres [6, 7, 8]. In supervised learning to predict prosody, there can also be a potential for variation between corpora due to different labelers, transcription conventions, and recording conditions. However, with the exception of [9, 10], there has been little previous work exploring the potential to train a classifier using labeled data from one domain or genre for deployment on another.

In this paper we conduct such experiments using the Boston University Radio News Corpus (BU-RNC) and a section of the Switchboard Corpus (SWBD) labeled with accents and prosodic phrase boundaries. The BU-RNC consists of read news stories by professional radio announcers, and has been widely used for research into automatic prosody classification. The SWBD Corpus consists of spontaneous telephone conversations between strangers on an assigned topic. We focus on two word-level binary classification tasks: presence vs. absence of pitch accent on a word, and presence vs. absence of an intonational phrase boundary (break) after a word. We use a large standard set of acoustic-prosodic features extracted from the speech waveform and transcripts, as well as a small set of textual features such as part-of-speech extracted from the transcripts only. We compare cross-genre classifier performance with acoustic and textual features, separately and together, finding that baseline performance is quite good; in most cases the best performance is achieved with both feature sets. We compare genre-specific training with multi-genre training, showing that most of the performance loss of cross-genre training is recovered by the multi-genre classifier; this suggests that decision boundaries in both genres can be modeled simultaneously. We then conduct unsupervised adaptation experiments based on the assumptions of sample selection bias, class proportion bias, and feature shift; none of the unsupervised adaptation techniques explored so far lead to consistent gains over the baseline cross-genre model.

2. Related Work

There has been some work comparing the utility of different feature sets for prosodic event detection in different genres. In [5], the authors explored the use of different textual features for predicting pitch accent in two corpora each of read speech (including BU-RNC) and spontaneous speech (including SWBD), finding that the best sets (for in-genre training) were slightly different but that a good universal set could be found that did almost as well. On the task of sentence and dialog act boundary detection, several papers by the group of Shriberg and Cuendet have compared feature utility and distributions across read news, meeting, and conversational speech genres. In [6], Cuendet et al. compared the relative usefulness of language and acoustic feature types for prediction in different genres, with and without the presence of speech recognition errors. In [7], acoustic-prosodic models were combined with language models for sentence segmentation, and their relative usefulness was compared on broadcast news and SWBD.

In our work, we do not compare utility of features for training and testing a classifier within each genre. Rather, we are interested in the problem of cross-genre training, as explored in [10] for pitch accents and [11] for sentence segmentation. Of most relevance here are [9, 11], which show success with supervised model adaptation between SWBD and a meetings corpus with a variety of methods, including model combination and using predictions of the cross domain model as a feature. The authors of [11] also conducted several experiments similar to ours. They compared feature types (acoustic-prosodic vs. n-grams) for cross-genre training between the meetings and a broadcast news corpus; prosodic features were better than n-grams on broadcast news, whether trained on meetings or news,

while n-grams were better than prosodic features on meetings only if trained on meetings. They also showed improved cross-classification performance in a "cheating experiment" in which they optimized the acceptance threshold on the test set, but were not able to approach the performance of within-genre training.

Our work differs from [9, 11] in several ways. First, their work considered only sentence segmentation, which is related to the phrase boundary task considered here, but is not equivalent. Second, they use a feature set consisting of n-grams and a smaller set of acoustic features. Third, their focus is supervised adaptation, using a small amount of labeled data from the target genre. Our goal in this work is to investigate the possibility of *unsupervised* adaptation. Our work is inspired partly by [8], which included a detailed comparison of acoustic feature distributions in a meeting corpus and a broadcast news corpus. They observed many similarities in the shapes of the distributions, and proposed that these similarities could be used for cross-genre training and adaptation; for instance, they suggested the possibility of feature normalization or automatic adjustment of class proportions. Their acoustic feature set is very similar to ours, although their task is dialog act segmentation, and they use different corpora.

3. Methods

The BU-RNC has been annotated using the Tones and Break Indices standard [12], which labels words on two tiers: the tone tier contains marks for the location of pitch accents and for distinguishing several accent categories characterized by patterns of high or low pitch, and the break tier consists of an index representing the strength of the boundary after each word. We follow a common approach of collapsing all pitch accent categories to a single "accent" label, resulting in a binary word classification problem (presence/absence of pitch accent). Similarly, for breaks, we map indices 4 and above (major phrase breaks) to a "break" class associated with the preceding word. A portion of the Switchboard I corpus has been labeled [13] with a variation of this standard: there are three pitch accent categories ("full", "weak", "none"), and boundaries after backchannels are given a distinct mark rather than a break index. We mapped both "full" and "weak" accent labels to the pitch accent category, and included the backchannel marks in the break category.

For these experiments we used a portion of BU-RNC drawn from six speakers, and a sampled subset of the same size from the labeled SWBD data, in order to control for size effects. Training sets were 13k words each, with no speakers shared between the train and test sets for either genre. The SWBD test set comprised both sides of four conversations, and the BU-RNC test set comprised the "labnews" section for three speakers.

Most of our features are fairly standard and have been used in other work. There are 64 acoustic features based on those described in [7], which represent pitch, energy, and duration information. They were derived from a phone-level forced alignment of the transcripts with the waveforms, and are associated with a word or the post-word boundary, with some representing differences between this word and the next, or in a window around the boundary. Duration features include word and pause durations as well as subword durations such as average/maximum/last vowel, last rhyme, and average/maximum phone. The same set of acoustic features was used for both the pitch accent and phrase break experiments. The textual features were extracted from the transcripts only, and include part-of-speech (POS) labels for the current, previous, and next word. These were derived from the MXPOST POS tagger [14] applied to sentences,[1] as identified in the BU-RNC transcripts, and to utterances, as identified in the MS-State SWBD transcripts.[2] For the pitch accent task, additional textual features include log unigram, bigram, and backwards bigram probability derived from a separate corpus of broadcast news and talk shows, and the *accent ratio* feature from [15], which represents the fraction of time a word occurs accented in the training set. For the break task, the textual feature set includes the three POS features plus two novel features: *break ratio* defined analogously to accent ratio, and *POS bigram break ratio*, defined analogously for the POS bigram formed from the current and next words.

4. Cross-Genre Training

In this section we present results using the BoosTexter [16] classifier, which was also used for the sentence segmentation experiments in [8, 6].[3] BoosTexter is based on the AdaBoost algorithm with single-feature decision stumps as the base classifiers; it easily incorporates both continuous and categorical features, and can handle missing feature values. Figure 1 shows the classification error rates for models trained on the in-genre training set, the cross-genre training set, and both sets concatenated together. (In the latter case, the "ratio" features were computed from the in-genre portions only).

We observe first that in most cases there is a significant increase in error from training on the cross-genre training set compared with in-genre training, but it is generally only a few percentage points. In one case—textual features only on BU-RNC breaks—the SWBD-trained classifier actually does significantly *better* than the in-genre classifier. However, the F measure is slightly lower—it classifies too few word boundaries as breaks, and gets a very low recall. The higher accuracy may be due to the fact that the BU-RNC train and test sets differ in class proportion (22% vs. 15% breaks); the BU-RNC classifier detects too many breaks, whereas the SWBD classifier detects too few, but gets better accuracy. We note that the SWBD train set is much more varied in speaker composition than the BU-RNC, whose training set is 70% composed of one speaker. Thus, SWBD may actually transfer better to new speakers for some feature sets.

We observe next that in most cases, the classifier trained with both genres together is able to achieve or nearly achieve the performance of the in-genre-only classifier. This suggests that in most cases, the classifier is able to model the decision boundaries for both genres simultaneously nearly as well as it can for each genre separately, although the results for BU-RNC breaks suggests there may be a small amount of mismatch in this case.

Comparing feature sets, we observe that with the exception of BU-RNC breaks, the cross-genre error increase appears more severe with textual features only than with acoustic or acoustic+textual. This seems to be due to a variety of factors: in some cases the POS features suffer substantial degradation, while in other cases the "ratio" features do not transfer well across genres. This is consistent with observed differences in [5] between read vs. spontaneous speech, in terms of the relationship between accented words and POS categories, and in terms of accent ratio distributions. Although textual features appear to

[1] We used a version of the tagger trained on Treebank-3.

[2] The MS-State transcripts are available at http://www.isip.piconepress.com/projects/switchboard/

[3] We used the open-source implementation icsiboost at: http://code.google.comp/icsiboost/

provide little additional benefit over acoustic features for cross-genre classification, performance is slightly better with both sets in all cases. Even in the case of SWBD breaks, where both feature sets together provide no benefit (actually slightly worsen performance) over textual features alone for the in-genre classifier, the cross-genre classifier benefits from using both feature sets. We conclude that both feature sets transfer well enough to be used in cross-genre classification, although some benefit might be achieved by feature selection within the subsets.

Our in-genre training results for pitch accent on BU-RNC are close to those reported by [10] using a different classifier and set of acoustic features, and our cross-genre results on BU-RNC using SWBD training are slightly better than their reported results using the BDC spontaneous speech corpus. Note also that SWBD appears to be a more challenging corpus than BDC, on which they are able to achieve 17% classification error rate. An interesting result in our case is that the classifiers trained on SWBD have lower error rates on the BU-RNC test set than on the SWBD test set.

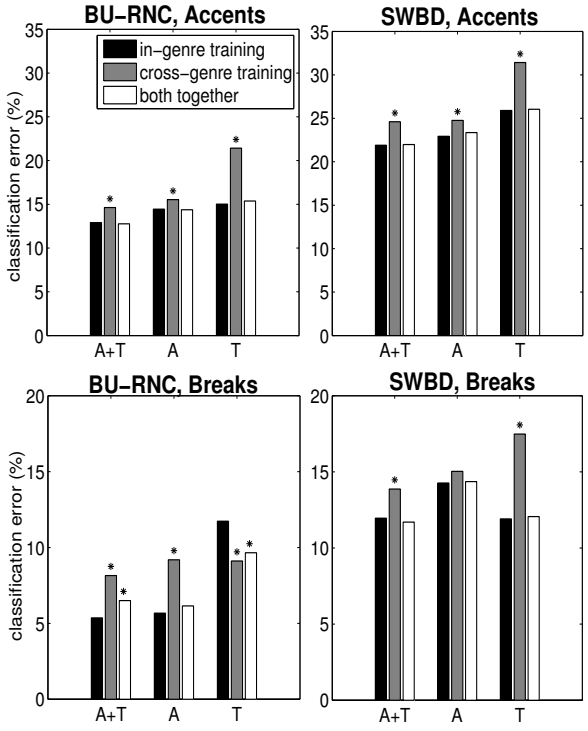

Figure 1: Baseline classification error rates for different feature sets ("A"=acoustic, "T"=textual features). The majority-class decision error rates are 45% (47%) for BU-RNC (SWBD) accents and 15% (22%) for BU-RNC (SWBD) breaks. Starred bars indicate that the classifier is significantly different from the corresponding within-genre training case (black) under McNemar's test ($p < 0.05$).

5. Adaptation Approaches

The goal of unsupervised domain adaptation is to use unlabeled test domain data along with labeled cross-domain training data to build a classifier tuned to the test domain. Several approaches have been suggested in the machine learning literature, which make various assumptions about the nature of the domain mismatch. One possible assumption is that of sampling selection bias: the domains have the same posterior label distribution $p(y|x)$ for label y given the features x, but the training domain examples represent a biased sample from the unconditional distribution $p(x)$; this suggests approaches such as instance weighting [17]. A different assumption is that domains share class generative distributions $p(x|y)$ but differ in class proportions $p(y)$; this suggests that automatic adjustment of the class priors may help, as in [18]. And yet another assumption is that features have been scaled or shifted between domains, which might be corrected by normalization. A goal of our experiments is to investigate whether unsupervised domain adaptation approaches based on these assumptions might be effective for accent and phrase boundary classification across genres.

Sampling selection bias In addition to the above experiments with the BoosTexter model, we performed similar experiments with a logistic regression (LR) model.[4] LR fits only a linear decision boundary, so it could be more sensitive to a sampling selection bias: if the decision boundary were not really linear and the training set represented a biased sample from the feature space $p(x)$ of the test domain, the learned solution maximizing likelihood on the training set might do poorly on the rest of the feature space [17]. BoosTexter can fit a more flexible decision boundary due to the decision stumps, which can split on arbitrary thresholds, so we believe it may be less prone to this type of mismatch. We did not observe a consistent benefit of one classifier over the other, with the exception that BoosTexter was generally better using textual features only, both within and across genres, likely because it handles categorical features better.

Using the LR model, we performed experiments with instance weighting to match the test set distribution, as described in [17]. In this approach, training samples are weighted higher in the learning objective when they come from high-density regions in the test domain. However, estimating the densities from samples is difficult in our setting, since we have high-dimensional feature vectors with both continuous and categorical features. Instead, we used a method inspired by [19], in which a logistic regression model is used to predict whether a given sample was generated by the train or test domain. The weight for each labeled training sample is taken as the odds of the sample belonging to the test domain. We tested this approach in the context of the LR-based prosody classifiers. In general it did not help, and in many cases actually worsened performance; one problem was that the training sample weights were mostly very small (by virtue of the samples being from the training domain), but a few were very large. We note that the LR classifier to predict domain membership achieved 83% accuracy on a separate test set. Perhaps the two genres do not substantially overlap in feature distribution, in which case little can be achieved by instance weighting (although the assumption of identical posterior distributions $p(y|x)$ may still be correct).

Class prior adjustment We expected some class frequency differences between genres. For instance, [5] showed that read speech has a higher proportion of words with pitch accents. We did not observe a large difference in the relative frequency of breaks between genres, although there were differences between the train and test sets for BU-RNC. For pitch accent we observed 56% occurrence in the BU-RNC training set, com-

[4] We used a modified version of the one at: http://www.informedia.cs.cmu.edu/yanrong/MATLABArsenal/MATLABArsenal.htm

pared with 41% in the SWBD training set (and 55% vs. 47% in the test sets). We tested whether changing the class prior $p(y)$ in the cross-genre model could improve performance. Using "cheating" experiments in which we adjusted the class probability outputs of the logistic regression classifier so that the prior term $p(y)$ matched the proportion seen in the test data, we observed slightly worse results for BU-RNC on SWBD, but a significant improvement in accuracy of 1% absolute for SWBD on BU-RNC. In the absence of known test set proportions, this approach can be implemented using iterative adjustment for $p(y)$ as suggested in [8, 18]; the unsupervised approach let to roughly the same improvement. Unfortunately, the unsupervised approach is unstable, since it depends on proportion estimates in the test data. When applied to other train/test pairs it sometimes led to worse performance: in the case of SWBD on BU-RNC breaks, which has a small initial classification proportion, the prior was adjusted down to 0%. In general, prior adjustment is not beneficial for all domain mismatch scenarios—even when adjusted in the correct direction, accuracy went down in some cases.

Feature normalization We tested normalization of the continuous acoustic features by subtracting the mean and dividing by the standard deviation in the corresponding genre's training set. This might help in the case that the train/test mismatch were mainly due to some constant shifting of the features, for instance, if the lower speaking rate in BU-RNC shifted duration features, making them all higher than those in SWBD. However, in general this did not work, and even worsened results significantly on the cross-genre accent task.

6. Conclusions

We have presented the results of cross-genre classification for two prosodic classification tasks and described some initial approaches to unsupervised domain adaptation. Baseline cross-genre classification is worse than in-genre classification, but still reasonably good in most cases. A small improvement was achieved by prior adjustment for the SWBD accent classifier tested on BU-RNC, but in other cases this approach led to worse results. None of the cross-genre classifiers were improved by sample selection bias correction or simple feature normalization. The fact that a general classifier can be built for both genres suggests that their conditional class distributions $p(y|x)$ may be similar, but the fact that a genre classifier achieves high training set separation suggests they may be mostly separated in the feature space. Future work might consider a more detailed feature selection approach, particularly within the acoustic features. This can be done automatically in an unsupervised fashion, by eliminating or penalizing features whose marginal distributions differ too much between genres. Another possibility is to apply semi-supervised learning methods such as bootstrapping or co-training, which make use of unlabeled data in training and have the ability to adapt to small changes in the decision boundary between train and test domains. Some success has been shown with these methods for in-genre training [20, 21].

Our experiments were based entirely on reference transcriptions and segmentations, but future work should consider features derived from speech recognition output, as considered in [11] for sentence segmentation. This could change conclusions significantly, for example if recognition quality differs greatly between domains.

7. References

[1] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 469–481, 1994.

[2] M. Hasegawa-Johnson et al., "Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus," *Speech Communication*, vol. 46, no. 3-4, pp. 418–439, July 2005.

[3] S. Ananthakrishnan and S. S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 216–228, 2008.

[4] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," Boston University, Tech. Rep., March 1995.

[5] J. Yuan, J. M. Brenier, and D. Jurafsky, "Pitch accent prediction: Effects of genre and speaker," in *Proc. Interspeech*, 2005.

[6] S. Cuendet et al., "Cross-genre feature comparisons for spoken sentence segmentation," in *Proc. Int. Conf. Semantic Computing*, 2007, pp. 265–274.

[7] E. Shriberg et al., "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, pp. 127–154, 2000.

[8] E. Shriberg et al., "Prosodic similarities of dialog act boundaries across speaking styles," in *Linguistic Patterns in Spontaneous Speech*, ser. Language and Linguistics Monograph Series A25, S. C. Tseng, Ed., 2009, pp. 213–239.

[9] S. Cuendet, D. Hakkani-Tür, and G. Tur, "Model adaptation for sentence segmentation from speech," in *IEEE Spoken Language Technology Workshop*, 2006, pp. 102–105.

[10] A. Rosenberg, "Automatic detection and classification of prosodic events," Ph.D. dissertation, Columbia University, 2009.

[11] S. Cuendet, "Model adaptation for sentence unit segmentation from speech," IDIAP Research Report 06-64, Tech. Rep., October 2006.

[12] K. Silverman et al., "ToBI: A standard for labeling English prosody," in *Proc. ICSLP*, 1992, pp. 867–870.

[13] S. Calhoun et al., "The NXT-format Switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," To Appear: Language Resources and Evaluation Journal.

[14] A. Ratnaparkhi, "A maximum entropy part-of-speech tagger," in *Proc. EMNLP*, 1996, pp. 133–141.

[15] A. Nenkova et al., "To memorize or to predict: Prominence labeling in conversational speech," in *Proc. HLT-NAACL*, 2007, pp. 9–16.

[16] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2, pp. 135–168, May 2000.

[17] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of Statistical Planning and Inference*, vol. 90, no. 2, pp. 227–244, October 2000.

[18] Y. S. Chan and H. T. Ng, "Estimating class priors in domain adaptation for word sense disambiguation," in *Proc. ACL*, 2006, pp. 89–96.

[19] S. Bickel, M. Brückner, and T. Scheffer, "Discriminative learning for differing training and test distributions," in *Proc. ICML*, 2007, pp. 81–88.

[20] J. H. Jeon and Y. Liu, "Semi-supervised learning for automatic prosodic event detection using co-training algorithm," in *Proc. ACL-IJCNLP*, August 2009, pp. 540–548.

[21] G.-A. Levow, "Unsupervised and semi-supervised learning of tone and pitch accent," in *Proc. HLT-NAACL*, 2006, pp. 224–231.