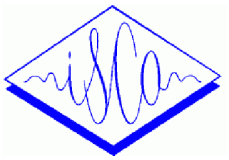


INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Interspeech Pathology Challenge: Investigations into Speaker and Sentence Specific Effects

Anthony Stark, Alireza Bayestehtashk, Meysam Asgari and Izhak Shafran

Center for Spoken Language Understanding, OHSU, Portland, OR
{starkan, bayesteh, asgari, shafrani}@ohsu.edu

Abstract

In this paper, we report our experiments on Interspeech 2012 Speaker Trait Pathology challenge task [2]. Specifically, we investigate two factors that impact the acoustic properties of the utterances collected in this task. Although the task treats utterances as independent data points, multiple utterances are recorded from individual speakers. Furthermore, the utterances correspond to readings of 17 given written sentences. In one experiment, we attempt to reduce variation due to speaker through dimensionality reduction. While these experiments showed promising results on development set, the performance did not translate to the evaluation test. In another, we learn classifiers conditioned on the sentences to capture sentence-specific signatures. This approach showed improved performance over the baseline on development set and the improvement translated to marginal gains on evaluation set. These experiments demonstrates the need to pay attention to the independence assumptions while collecting and defining clinical tasks.

Index Terms: speech pathology, intelligibility

10.21437/Interspeech.2012-104

1. Introduction

In this paper, we detail an experimental analysis of the Interspeech 2012 speaker trait challenge. Specifically, we are interested in the pathology sub-challenge; a problem tasked with discriminating healthy voice samples from those influenced by the presence of throat cancers. Though of interest in its own right, the general problem of characterizing pathologies via spoken language opens a wide range of diagnostic possibilities. Given an ability to discriminate such pathologies via language could provide valuable screening tools in the future.

The challenge itself is concerned with the effects of certain cancers on voice intelligibility. To this end, the challenge is a binary classification problem with *intelligible* (I) and *non-intelligible* (NI) labels corresponding to the underlying pathology. Subjects, both controls and patients, were asked to read 17 sentences and they were recorded in separate files. For conciseness, we refer the reader to the challenge paper [2] for detail regarding corpus collection and preparation. The rest of this paper is arranged as follows. In Section 2, we detail the characteristics of the corpus, the baseline features provided, and our experimental efforts applied to them. We then analyze the effect of dimension reduction on the corpora in order to better improve generalization of our classifiers. In Section 3, we extend our analysis beyond the baseline acoustic feature set to include sentence-conditioned models. Finally, in Section 4 we wrap up the paper with some concluding remarks and ideas for future direction.

2. Speaker Specific Effects on Classifier

In this section we study the classification problem by manipulating the acoustic feature vectors provided by the challenge paper [2]. The provided corpus consists of of approximately 6000 acoustic features for each of the 2386 audio samples. Despite the high dimension, the feature itself is derived from a relatively small pool of acoustic phenomena (pitch, energy, MFCC, etc) and their statistics (min, max, quantiles, etc). The corpus includes metadata about the age, the gender and the regional dialect of the speaker, in addition to the I/NI (intelligible / non-intelligible) labels assigned by independent listeners.

The corpus treats sentences read by the same speaker as independent samples, which is clearly not the case. We expect a large degree of mutual information between the identity of the speaker and the pathology of their samples – a healthy person nominally producing intelligible samples and vice versa. This poses a problem in learning a classifier, since most classifiers treat samples to be independent. A classifier that ignores the dependence will inadvertently be recognizing speakers instead of the pathological conditions (the assigned labels). Such a classifier is unlikely to generalize to new data.

One way to alleviate this problem is to tune the classifier on a cross-validation set, where the partitions are created in a manner that does not permit utterances from the same speaker to be represented in different folds (Table 1, CV-B). Though the challenge does not explicitly provide speaker labels, we have attempted to recreate our own using a tuple of the age, gender and I/NI labels provided. Though this does not preclude the possibility of labeling multiple individuals as the same person, it does allow us to construct partitions with non-overlapping speakers.

Table 1: *Influence of speaker-specific factors on the classification. CV-B partitions the data ignoring the speaker overlap between folds, CV-B partitions the data at the speakers-level. Accuracy from linear SVM is reported as class weighted averages.*

Experiment	Accuracy
CV-A Full	85.5
CV-B Full	59.0
CV-A PCA-500	86.2
CV-A PCA-200	84.6
CV-A PCA-100	81.8
CV-A PCA-50	79.9
CV-A PCA-10	69.3
CV-B PCA-500	59.3
CV-B PCA-200	59.3
CV-B PCA-100	59.1
CV-B PCA-50	57.9
CV-B PCA-10	58.1

For comparison, we also create a cross-validation set by randomly partitioning the utterances ignoring the dependence between samples (Table 1 CV-A). Using both schemes, we test the original feature vector as well as vectors projected onto lower dimension principal component analysis (PCA) learned on the global data. We learn two linear SVMs whose parameters are optimized using the two cross-validation sets. The classifiers were learned using libSVM [1]. The performance – class weighted recognition accuracy – of the two classifiers are reported in Table 1. For the full 6k feature vector (CV-A Full and CV-B Full) it is immediately apparent the linear SVM models are learning speaker-specific traits. The difference between the two CV schemes indicate that the linear SVM is far more adept at learning a speaker-specific discriminator, rather than a general intelligibility discriminator.

The first option that comes to mind to address this issue is a form of dimensionality reduction. For this application, PCA is particularly appealing as the 6k feature dimension is already known to be derived from a relatively small number of acoustic features. Interestingly, the PCA reduction rapidly diminishes the ability of the random-CV (CV-A) classifier in exploiting the inappropriate speaker information. Using the full subspace, speaker specific data (CV-A) gives approximately 25% absolute improvement over the speaker-separated classifier (CV-B). These false gains are shrunk considerably with application of PCA reduction. Despite this, PCA doesn't appear to yield improved generalization, as the more principled speaker-segmented CV (CV-B) showed relatively little sensitivity to the dimension reduction. However, given no degradation was observed, it does provide a mechanism for exploring classifiers suited for lower dimension problems.

Table 2: *Weighted classification accuracy of linear and RBF SVMs for the PCA-reduced feature vectors.*

Experiment	Dev. Accuracy
LIN-500	61.0
LIN-200	61.1
LIN-100	61.0
LIN-50	60.7
LIN-25	61.9
LIN-10	59.9
RBF-500	64.6
RBF-200	66.1
RBF-100	65.4
RBF-50	66.9
RBF-25	66.1
RBF-10	64.5

The two classes may no longer be easily separable by a linear classifier and so we learn an SVM with radial basis function (RBF), which has been successfully employed in learning non-linear boundaries. The results reported in Table 2 show significant improvement over linear SVM, as we expected. Cross-validation folds (without speaker overlap) were used to tune SVM parameters, with a final model being learned on the full training set.

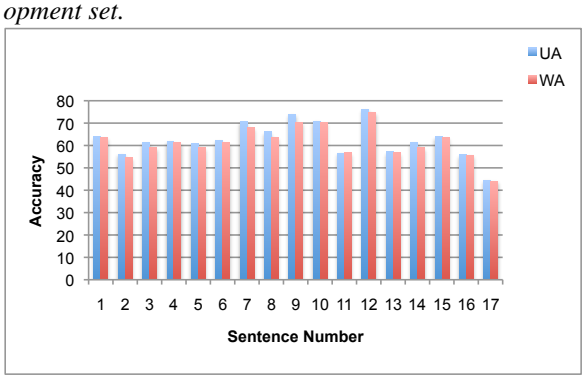
The baseline linear SVM results (WA) for the development set [2] are 61.3% – giving the PCA linear SVM (61.94% accuracy) marginal advantage at best. The RBF models however, showed promising gains with unweighted average recall reaching 66.89% on the development set. This was also higher than non-linear random forest classifier given in the baseline system (65.1%) [2]. The fact that the baseline experiments performed relatively poorly with a linear classifier suggest the problem is not linearly separable despite its large dimension.

For the test set, all labeled data was pooled for training and a new model learned using parameters found with cross validation. The RBF models saw consistent performance across the CV-training, development and final testing corpora. Despite the consistency, the class weighted and unweighted accuracy on evaluation set was found to be 64.14 and 66.97 respectively, which was lower than that of the reported baseline system [2]. This is most likely due to the fact that the labels in the evaluation were skewed (64 vs 36) with respect to the training and development set which had approximately balanced class labels.

3. Sentence Specific Effects on Classifier

As mentioned earlier, the utterances correspond to 17 sentences that were read by the subjects. As the utter-

Figure 1: *Per-sentence classification accuracy (unweighted) of sentence-conditioned linear SVMs on development set.*



ances themselves are relatively short, we expect acoustic content to vary considerably across utterances corresponding to different sentences. The models considered in the previous experiments, as well as those appearing in the challenge baseline [2] ignore this source of variability.

In this section, we examine the effect of conditioning the classifier on the sentences. We partition the corpus into 17 separate, sentence-conditioned training, development and testing subsets. We train and test separate classifiers for each. This allows models to concentrate on learning an N/NI discrimination rather than the less important sentence-to-sentence variability. Of course, this approach also yields less training data for each classifier and so we learned a simple linear SVM for each subset. Incidentally, this also allows us to directly compare the results with that of linear SVM reported in the baseline system [2].

The accuracies of the sentenced conditioned classifier on development and test set are reported in Table 3 and illustrated in Figure 1 for easy interpretation. The results for each sentence conditioned classifier show significant spread in accuracy, suggesting the choice of underlying text is an important consideration for the pathology task. However, we have not yet been able to reliably determine whether the proportion of voiced/unvoiced speech, phonetic content, utterance length or other some other factor is in play. While overall classification showed marginal improvement over the baseline, it is interesting to note that like the baseline linear classifier, there is an unexplained improvement in performance when evaluating over the test set. Overall the sentence conditioned models yielded a 0.8% unweighted average recall improvement over the baseline random forest classifier.

4. Conclusions

In this paper, we explored two sources of variability in the corpus collected for the Interspeech 2012 pathology challenge. While the utterances from speakers reading 17 sentences are treated as independent samples, they are not. In one set of experiments, we attempt to alleviate the speaker overlap by learning a classifier on a 20-fold cross-validation set, where the folds were carefully chosen to prohibit utterances from same speakers appear in multiple folds. We compare the performance with randomly partitioned 20-fold cross-validation set. Our results on development set were consistent with the expectation that eliminating speaker overlap across the fold improves performance. When the features were reduced using PCA and a non-linear boundary was learned using RBF kernels, we obtained better performance than from random forest in the baseline system [2]. However, due to the skew of the labels in the test set, the improvements did not translate to the evaluation set. In another thread of experiments, we investigated the effect of conditioning the classifier on the sentence by training separate classifiers for each sentence. The resulting classifier performed better on both development and evaluation set. We plan to analyze our classifiers in more detail once the evaluation set is released.

Table 3: *Development set classification accuracy of sentence conditioned linear SVMs. Unweighted average recall and weighted average recall (given in brackets).*

Sentence	Accuracy
”Hij at van een gouden...”	63.9 (63.6)
”En toen hij achttien jaar...”	56.2 (54.5)
”De prins stond bij het raam”	61.4 (59.0)
”Toen hij nog in de wieg lag”	61.8 (61.3)
”Zij hadden ieder een...”	60.8 (59.0)
”De prins was erg verwend”	62.2 (61.3)
”Maar hij was toch jarig...”	70.7 (68.1)
”Had hij alles wat hij...”	66.1 (63.6)
”maar ze waren erg verlegen”	74.0 (70.4)
”Al zijn speelgoed was van...”	70.9 (70.4)
”wat hij al niet had”	56.6 (56.8)
”want ze begrepen wel dat...”	76.1 (75.0)
”kreeg hij al een gouden...”	57.2 (56.8)
”toen zijn ooms en tantes...”	61.4 (59.0)
”en het werd steeds moeilijker...”	64.1 (63.6)
”Er leefden eens een koning...”	55.9 (55.8)
”Dat was de prins”	44.4 (44.1)
Total development set	61.3 (62.4)
Total test set	72.1 (70.4)

5. Acknowledgements

This research was supported in part by NIH awards 5K25AG033723, 5R01AG027481 and P30 AG024978-05 and NSF awards 1027834, 0958585, and 0905095. Any opinions, findings, conclusions or recommendations

ex- pressed in this publication are those of the authors and do not reflect the views of the NIH or NSF.

6. References

[1] C. Chang, C. Lin, "LIBSVM A library for support vector machines", ACM Transactions on Intelligent Systems and Technology, vol.2, 27:1–27:27, 2011.

[2] B. Schuller, S. Steidl, A. Batliner, E.Noth, A. Vinciarelli, F. Buckhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, B. Weiss: "The INTERSPEECH 2012 Speaker Trait Challenge ", INTERSPEECH 2012.

[3] J.O. de Lira, K.Z. Ortiz, A. Carvalho Campanha, P.H.F. Bertolucci, and T.S.C. Minett. Microlinguistic aspects of the oral narrative in patients with Alzheimers disease. International Psychogeriatrics, FirstView:19, 2010.