Text-dependent pathological voice detection

Gopala Krishna Anumanchipalli[†‡], *Hugo Meinedo*[‡], *Miguel Bugalho*[‡],
Isabel Trancoso[‡], *Luís C. Oliveira*[‡], *Alan W Black*[†]

[†]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[‡]Spoken Language Systems Laboratory, INESC-ID/IST Lisboa, Portugal
{gopalakr,awb}@cs.cmu.edu {meinedo,mmfb,imt,lco}@inesc-id.pt

Abstract

While global characteristics of the speaker's source and spectral features have been successfully employed in pathological voice detection, the underlying text has largely been ignored. In this work, we focus on experiments that exploit the text stimulus that is read by the subject. Features derived from text include the mean cepstral distortion of the subject from an average intelligible speaker, and prosodic features include the speaking rate, statistics of phoneme durations, etc. The phonetic labeling information is also exploited to ignore all the unvoiced regions of the speech samples to improve the discriminability between intelligible and pathological voices. We also designed features that capture the speaker's overall closeness to intelligible instances of the same text stimulus from other speakers. Our experiments show that the proposed text-derived features improve the detection of pathological voices by 20%.

Index Terms: Pathological voices, example based detection, text-driven features, fusion of classification methods.

1. Introduction

Research on automatic extraction of paralinguistic information has rapidly grown in the last years, motivated by their importance in fields like Human-Machine interaction, Multimedia Retrieval, or in the educational and medical domains. The INTERSPEECH 2012 Speaker Trait Challenge [1] addresses three important speaker traits: personality, likability and pathology in speech. Like in previous challenges, the goal of this competition is to bridge the gap between research in this area and low compatibility of results, which is even more important for these traits given their highly subjective nature.

In this paper we will describe two contributions: a set of features that take advantage of the underlying text stimulus for pathological voice detection and a framework based on the calibration and fusion of several classifiers. Although our work is focused mainly on the pathology sub-challenge, given the proposed novel features, the framework was also applied to the other subtasks, and results for all sub-challenges will be presented.

Speech is the most important form of direct communication and any pathology that affects the speaking capabilities will have a large impact both on the subject's professional as well as social activities. Intelligibility assessment of pathological voices can be very relevant both for diagnostic and therapy evaluation. The assessment of voice quality can be made by a diagnostician or by direct examination (for instance with laryngostroboscopy). Hakkesteegt [2] evaluates two methods for the assessment of voice quality: Dysphonia Severity Index (DSI) and Voice Handicap Index (VHI). While the first is a combination of measurements that can be retrieved from recorded speech, in fact some are actually present in the baseline features of the challenge (like F0 or Jitter), the second is based on a questionnaire, and is therefor not suitable for automatic evaluation. In the work of Silva et al. [3] different methods of jitter estimation were used to detect larynx pathologies that can affect speech, such as vocal fold nodules or a vocal fold polyp. Our approach also uses an enlarged set of acoustic and prosodic features such as the described above but most importantly explores information from the phonetic content which has been largely under used and shows how the fusion of several distinct methods can be used to increase the performance of a pathological voice assessment system.

The paper is organized as follows: in Section 2 we present the novel features developed for the Pathology sub-challenge that take into account the underlying text stimulus. In Section 3 we review the acoustic and frame level features used in Pathology and also in the other two sub-challenges. In Section 4 we present the different classification methods that were used in this work. Section 5 presents the results obtained by the developed methods in the three sub-challenges. Finally we draw some conclusions in Section 6.

2. Features exploiting the underlying text stimulus

Since we intend to exploit the underlying text stimulus that was read by the subjects, we device features that for which the extraction is made possible only due to the availability of the text. These features, that are both spectral and prosodic in nature, are explained in greater detail below.

As a preprocessing step, the entire corpus was phonetically aligned. Though the corpus provided manual phonetic segmentation, for precise consistency of the phone boundaries, we align the corpus automatically using the `ehmm` utility in festvox [4]. Since not all speech samples are intelligible, we train the phonetic models only on the intelligible examples from the training data, but align all instances using the same trained models. Additionally, given the phonetic alignments on all the intelligible utterances, a statistical parametric speech synthesis model [5] is built for the task. A Festival frontend for Dutch is built to process each sentence and automatically convert it into a string of appropriate Dutch phonemes. The idea here was to capture the characteristics of an intelligible Dutch speaker (based on all the intelligible utterances in the corpus) and to use them as a reference to test any new speech instance. More information on training statistical synthetic voices may be found in [6].

2.1. Mel-Cepstral Distortion

Using the statistical spectral models [5] built only from the intelligible speakers, it is possible to compute a 'distance' between an utterance and a synthetic utterance generated from the models. While DTW techniques can be used to compute the distance between the original and synthesized waveforms, usually the same durations as those employed by the speaker are used for the synthetic utterance so that there aren't errors due to sub-optimal alignments. This way, an accurate frame level error can be computed between the speaker and the estimate of spectra from the statistical model.

$$\mathrm{M}CD = 10/ln10\sqrt{s\sum_{d=1}^{i=24}(mc_d^{(t)} - mc_d^{(p)})^2}$$

We use the Mel-Cepstral distortion measure given by the above equation to compute the distance between mel cepstra of the speaker and the estimate, $mc_d^{(t)}$ and $mc_d^{(e)}$ respectively. We also exclude pause regions from this computation as they are irrelevant to this measure.

2.2. Speech rate

Speech rate may be defined as the number of syllable nuclei uttered per second, ignoring the utterance initial and final pauses. Speech rate has often been used as a feature in fluency studies and it may help capturing some characteristics of voice pathologies. This feature could be computed without the need for a transcription using for instance the Praat script [7], but the availability of the automatically aligned utterances made it handy to compute speech rate in this alternative way.

2.3. Phoneme duration prediction error

The default statistical duration model in Clustergen is a decision tree containing linguistic questions about the context of the current phoneme, where the leaf nodes contain the average duration of the training instances that fall in that path. Given the phonetic segmentation, one can hence compute a mean prediction error per phoneme for each utterance. This feature captures some information about any abnormal durations which may be relevant for pathology detection.

2.4. F0 prediction error

Similarly to what has been done for duration, we also include the F0 prediction error for each utterance. This however may be much less informative, as intelligible speakers that speak emphatically may also be 'far' from the average predictable F0.

2.5. Distance from positive examples of the stimulus

A particular characteristic of this corpusis that there were only 17 sentences which each speaker had recorded. This lets us devise features that capture the distance of an instance from all the intelligible exemplars of that text stimulus uttered by other speakers. Though several measures can capture this distance, we use the L2 distance of the default smile feature set of the utterance from the average vector for intelligible speakers.

Table 1 presents the correlation of each of the above features with class labels and intelligibility over unseen data. It can be seen that the features provide significant correlations, and the trends are also reasonable. It is accurate that MCD, and phone duration error are negatively correlated with intelligibility. It is interesting that F0 prediction error is not negatively correlated. This means that intelligible speakers have unique speaking styles (by being emphatic etc.,) and being far away from the mean model is not essentially bad. The only confounding case is the distance from positive examples, which is positively correlated with intelligibility. This could be an artefact of the small size of the corpus, or the fact that the script and transcript are very different in many cases.

Table 1: Correlation of each feature with the Pathology class/intelligibility on unseen data

Feature from text	Correlation	
	Boolean Classification	Intelligibility index
MCD	-0.17	-0.17
Speech rate	0.298	0.364
Phoneme duration prediction error	-0.151	-0.114
F0 prediction error	0.117	0.142
Distance from positive examples	0.221	0.175

3. Acoustic Features

We used the baseline acoustic feature set provided by the challenge organizers [1] which consists of a vector with 6125 features obtained for each segment (referred as "arff6125" in this work). Additionally we modified the openSMILE feature extractor [8] configuration file in order to retain only the frame based low-level descriptors (LLD) and their deltas. This consists of a vector with 128 coefficients obtained every 10ms and referred as "arff128" in this work. We also investigated how more traditional cepstral frame level features such as Perceptual Linear Prediction (PLP) [9] behave in these three challenges. Although these cepstral features were originally developed for speech recognition they have obtained surprisingly good results in many audio and speech classification tasks. This front-end, here referred as "PLP26", extracts from the audio every 10ms a frame with 12th order PLP coefficients plus energy plus deltas.

For the frame level feature extraction methods the non-speech portions longer than 0.2 seconds were removed in an effort to improve discriminability by extracting features only during speech (voiced) segments.

4. Classification Methods

In this work we used different classification methods, starting by reproducing the methods used in the competition baselines [1]. For this purpose we used the open-source classifier implementations from the WEKA data mining toolkit [10]. First we evaluated linear Support Vector Machines (SVM) trained with Sequential Minimal Optimization (SMO), as they are robust against overfitting in high dimensional feature spaces. A slightly larger range has been tested for the parameter C $\{10^{-6}, 10^{-5}, \ldots, 10^{+3}\}$, in the SMO experiments, with best results at 10^{-1} in the devel set. Secondly, we evaluated Random Forests (RF), which avoid the curse of dimensionality by constructing ensembles of REPTrees trained on random feature subspaces [1].

4.1. Baseline methods with novel features

We also used the baseline methods, SMO and Random Forests (RF) to obtain classifications for the five novel features previously described. Again for the SMO we tested a slightly larger range for the parameter C $\{10^{-6}, 10^{-5}, \ldots, 10^{+3}\}$, with best results at 10^{-1}. For the Random Forests we used the same range for the parameters, with best results of 500 for the number of trees and 0.02 for sub-space size. As described in the challenge article [1], we also used the training set to train the models and the development set to tune the parameters. For the test set, we also retrained the models using both train and devel sets with the previous parameters.

4.2. SVM with ARFF128 features

The "ARFF128" frame level features served as input to a linear kernel Support Vector Machines, for which we used the LibSVM toolkit [11] with best value for parameter C as 10^{-3} estimated in the devel set.

4.3. MLP with PLP26 features

For the "PLP26" features we used a Multi-Layer Perceptron classification paradigm, trained with our own simulator [12]. This MLP takes as input context 21 contiguous frames of features and has two hidden layers of 150 and 100 units. This configuration of hidden units was the one that achieved the better classification scores in the devel set.

4.4. Calibration and Fusion Back-End

Linear logistic regression fusion and calibration of the developed front-end systems has been done with the FoCal Multiclass Toolkit [13]. The output log-likelihood ratio (llr) scores from this fusion back-end were later converted into probabilities, which is more meaningful in terms of human analysis. This was achieved by scaling the scores to produce confidence values with the expression (1).

$$p(score(t)) = \frac{e^{score(t)}}{\sum_k e^{score(k)}} \quad (1)$$

Several experiments of fusing the different front-ends were tested. The ones that obtained better results are presented in Table 5.1, Table 5.2, Table 5.3 and discussed in the next section.

5. Experimental Setup and Results

Results are expressed in terms of Unweighted and Weighted Accuracy on average per class (% UA and %WA). The former (%UA) is the relevant measure for the competition since it compensates when the distribution among different classes is not well balanced [1].

5.1. Pathology sub-challenge results

In the Pathology sub-challenge, the devel set which has 746 segments is almost as large as the test set, but has the opposite balance (more non-intelligible segments, whereas the test set has many more intelligible segments). Since the methods we use do not need such a large set to tune classification and fusion parameters, we propose a subdivision of the devel set which is two-fold: first, allows for parameter tuning in a still rather large set and second, allows us to evaluate the performance of the developed systems in a separate set (from tuning) which was chosen to have a class balance very similar the test set. The two subsets derived from this new partitioning of the devel set were named "dev-tune" and "dev-test".Table 2 summarizes the number of segments for each class of the new devel subsets.

Table 2: *Partitioning of the devel set.*

Class	dev-tune	dev-test	devel (total)
I	251	90	341
NI	355	50	405
Total	606	140	746

Table 5.1 summarizes the results obtained in the development set by the different systems individually and by their combination using the calibration and llr fusion back-end.

Table 3: *Pathology sub-challenge results.*

Systems (% UA)		dev-tune	dev-test	devel
a	SMO - ARFF6125	61.1	61.4	61.4
b	RF - ARFF6125	62.5	64.1	62.7
c	CART - ARFF6125	66.3	63.0	65.7
d	SVM - ARFF128	51.0	51.4	51.1
e	MLP - PLP26	53.4	48.9	52.5
f	SMO - 5-TXT-FEATS	82.2	81.3	82.2
g	RF - 5-TXT-FEATS	73.6	75.2	74.1
h	Fusion - a + b + d + f + g	—	81.9	82.6

The systems that use the new developed text features are very promising having obtained the best unweighted accuracy results (82.2%) when compared with all other acoustic features. System "f" represents an improvement of 20.8% absolute when compared with the competition baseline using the same classification method ("a"). Furthermore, the fusion which combines systems "a", "b", "d", "f" and "g" obtained a slightly better result than the best individual system ("f"). We chose these five systems for the fusion because the remaining two, "c" and "e", exhibit a much lower UA% when evaluated in the dev-test subset. This could mean that these systems will not generalize correctly when evaluated in other test sets. The fusion of these five systems represents an improvement of 17.5% absolute over the best competition baseline individual system which obtained 65.1% in the devel set [1].

The relative improvement is also seen on the actual challenge test set, where the Fusion system received a 66.3% unweighted recall against 61.59% of the Fusion system without the textual features. However this is still lesser than the challenge baseline of 68.9% unweighted accuracy. This is perhaps due to the different kinds of pathologies in the development and the test sets. Also the fact that the training data set doesn't have enough intelligible speakers to train a reliable 'average' intelligible speaker. We hope to evaluate on more speakers in training data and build gender specific and pathology specific models for better generalization of these features on unseen test speakers.

5.2. Likability sub-challenge results

In the Likability sub-challenge, the new text dependent features introduced for the Pathology sub-challenge could not be used, since there is no phonetic alignment. In a similar fashion to the Pathology sub-challenge, we reproduced the baseline methods and additionally used acoustic features (ARFF128 and PLP26) with different classification paradigms. Table 5.2 summarizes the results obtained in the devel set by the different systems

individually and by their combination using the calibration and llr fusion back-end.

Table 4: *Likability sub-challenge Results.*

Systems (% UA)		devel
a	SMO - ARFF6125	59.1
b	RF - ARFF6125	58.8
c	SVM - ARFF128	52.6
d	MLP - PLP26	51.3
e	Fusion - a + b + c + d	59.4

In this case, the fusion of individual systems produced only slight improvements over the competition baseline (0.9% absolute improvement over the baseline SMO method that had 58.5% UA). On the actual challenge test set, this system provided UA of 53.95%.

5.3. Personality sub-challenge results

In a similar fashion to the Pathology sub-challenge we reproduced the baseline methods and additionally used acoustic features (ARFF128 and PLP26) with different classification paradigms. Table 5.3 summarizes the results obtained in the devel set by the different systems individually and by the combination of them using the calibration and llr fusion back-end.

Table 5: *Personality sub-challenge Results.*

Systems (Mean % UA)		devel
a	SMO - ARFF6125	70.3
b	RF - ARFF6125	70.3
c	SVM - ARFF128	65.4
d	MLP - PLP26	64.3
e	Fusion - a + b + c + d	70.7

Again the fusion of individual systems produced a slightly better result when compared with the competition baseline, in this case achieving 0.6% absolute improvement over the baseline SMO method that had 70.3% UA. The UA on the actual challenge test set are as shown in Table 5.3. These are very comparable and in some classes (classes C, E and N) better than the challenge baseline results. The average UA on all classes we obtain is 68.14% and the challenge baseline is at 68.3%.

class	O	C	E	A	N	mean
UA	57.84	80.11	75.76	60.68	66.32	68.14

6. Conclusions

This work exploits the fact that the pathological voice recordings were done using read speech, thus allowing the use of the underlying text. While global characteristics of the speaker's source and spectral features have been successfully employed in pathological voice detection, the underlying text has been largely ignored. Features derived from text include the mean cepstral distortion of the subject from an average intelligible speaker, and prosodic features include the speaking rate, statistics of phoneme durations, etc.. We also adopted features that capture the speaker's overall closeness to intelligible instances of the same text stimulus from other speakers. Our experiments show that the proposed text-derived features improve the detection of pathological voices by 20% when compared with the competition baseline. This improvement, however, is obtained at the cost of introducing a language dependency in the detection of pathological voices. It would be interesting to further pursuit this study with an extended corpus, and in particular investigate the role of the different text-based features.

7. Acknowledgements

This work is supported partly by the Fundação de Ciência e Tecnologia through the joint program between the Portuguese Government and Carnegie Mellon University. This work was supported by FCT (INESC-ID multi-annual funding) through the PIDDAC Program funds.

8. References

[1] B. Schuller, S. Steidl, A. Batliner, E. Noeth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The interspeech 2012 speaker trait challenge," in *Proc. Interspeech 2012*, Portland, OR, USA, 2012.

[2] M. Hakkesteegt, *Evaluation of Voice Disorders: Dysphonia Severity Index and Voice Handicap Index*. Erasmus University Rotterdam, 2009.

[3] D. Silva, L. Oliveira, and M. Andrea, "Jitter estimation algorithms for detection of pathological voices," *EURASIP Journal on Advances in Signal Processing*, pp. 1–9, 2009.

[4] G. K. Anumanchipalli, K. Prahallad, and A. W. Black, "Festvox: Tools for Creation and Analyses of Large Speech Corpora ," in *Workshop on Very Large Scale Phonetics Research*, UPenn, Philadelphia, January 2011.

[5] A. Black, "Clustergen: A statistical parametric synthesizer using trajectory modeling," in *Proc. Interspeech 2006*, Pittsburgh, PA, 2006.

[6] A. Black and K. Lenzo, "Building voices in the Festival speech synthesis system," 2000, http://festvox.org/bsv/.

[7] N. Jong and T. Wempe, *Praat script to detect syllable nuclei and measure speech rate automatically*. Behavior research methods, 41 (2), 385 - 390, 2009.

[8] F. Eyben, M. Woellmer, and B. Schuller, "openEAR - introducing the munich open-source emotion and affect recognition toolkit," in *Proc. ACII 2009*, Amsterdam, Holland, 2009.

[9] H. Hermansky, N. Morgan, A. Baya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Proc. ICASSP 1992*, San Francisco, USA, 1992.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, , and I. Witten, "The weka data mining software: An update," in *SIGKDD Explorations*, vol. vol. 11, 2009.

[11] C.-C. Chang and C.-J. Lin, "Libsvm - a library for support vector machines," http://www.csie.ntu.edu.tw/cjlin/libsvm/index.html.

[12] H. Meinedo, "Audio pre-processing and speech recognition for broadcast news," Ph.D. dissertation, IST, Lisboa, Portugal, 2008.

[13] N. Brummer, "Focal multiclass toolkit," http://niko.brummer.googlepages.com/focalmulticlass.