The Thirty-Second International Florida
Artificial Intelligence Research Society Conference (FLAIRS-32)

Classification of Spontaneous Speech of Individuals with Dementia Based on Automatic Prosody Analysis Using Support Vector Machines (SVM)

Roelant Ossewaarde,[1,2] **Roel Jonkers,**[1] **Fedor Jalvingh,**[1,3] **Roelien Bastiaanse**[1,4]
[1]Center for Language and Cognition Groningen (CLCG), Rijksuniversiteit Groningen, Netherlands
[2]HU University of Applied Science, Institute for ICT, Utrecht, Netherlands
[3]St. Marienhospital - Vechta, Geriatric Clinic Vechta, Germany
[4]Center for Language and Brain, Higher School of Economics, Moscow, Russia

Abstract

Analysis of spontaneous speech is an important tool for clinical linguists to diagnose various types of neurodegenerative disease that affect the language processing areas. Prosody, fluency and voice quality may be affected in individuals with Parkinson's disease (PD, degradation of voice quality, unstable pitch), Alzheimer's disease (AD, monotonic pitch), and the non-fluent type of Primary Progressive Aphasia (PPA-NF, hesitant, non-fluent speech). In this study, the performance of a SVM classifier is evaluated that is trained on acoustic features only. The goal is to distinguish different types of brain damage based on recorded speech. Results show that the classifier can distinguish some dementia types (PPA-NF, AD), but not others (PD).

Introduction

Aphasia is an impairment to understand or produce speech as a result of brain damage, for example as caused by dementia. One of the aspects of speech that may be affected is prosody. In clinical practice, the transcription and analysis of connected speech of speakers with aphasia is an important diagnostic tool, but also time consuming and error prone. Automating language transcription is difficult because Automatic Speech Recognition (ASR) performance degrades significantly when applied to speech of speakers with aphasia, due to the difference between speech and language use of speakers without and speakers with aphasia.

Some types of speech and language disorders disrupt specific regularities in prosody. For example, the ability to maintain constant voicing or vary pitch height is affected in persons with Parkinson's Disease (PD, *e.g.*, Forrest, Weismer, and Turner 1989). Alzheimer's Disease (AD) sometimes causes speakers to produce shorter sentences and may influence the range of pitch variations (Kato et al. 2013). The non-fluent form of Primary Progressive Aphasia (PPA, Gorno-Tempini et al. 2004) - a form of progressive damage to the language areas of the brain - can disrupt the ability to speak fluently. A second type of PPA, semantic dementia, affects (morpho-)syntactic processing, but its clinical definition does not include problems with prosody.

Some features of pathological speech are easily detectable by human listeners. Recent studies show that human accuracy can be approached by machine learning models. For example, *inter alia*, Roark et al. (2011) have shown success at detecting various dementia types in English speaking participants with machine learning models trained on a combination of prosodic and syntactic variables.

Prosodic variables are the easiest to obtain, because they can be computed straight from the acoustic signal, independent of any parsing or morphosyntactic analysis. However, it is unclear how well models still perform if limited to only prosody, and studies on languages other than English are still scarce.

In the context of a larger study of processing of verbs and nouns in speakers with different types of dementia, currently performed by one of the co-authors (FJ), a corpus of connected speech from German speakers was created that includes speech from speakers with various forms of dementia. The aim of the current study is to determine the accuracy of a machine learned model to predict dementia based on only prosodic variables in this German corpus.

Methods

Speech data of a total of nine spontaneous conversations at three different moments in time were analyzed, with participants from different groups: non-brain-damaged individuals (NBD, n=7 participants) and participants with a clinical diagnosis of a form of neurodegenerative disease: probable Alzheimer's disease (AD, n=9), non-fluent primary progressive aphasia (PPA-NF, n=2), semantic dementia (PPA-SD, n=1), Parkinson's disease with MCI or dementia (PD, n=6). The average conversation length was 5m47s ($\pm$ 2m30s).

Common characteristics of speech are fluency, voice quality and prosody (Baken and Orlikoff 2000), in this study operationalized as frequency and length of pauses (fluency), mean Cepstral Peak Prominence (voice quality) and degree of pitch trend deviation (prosody).

The 22 hours of speech were automatically analyzed for speech and pauses using our own R-implementation of the Voice Activity Detection algorithm (VAD) based on the proposal by Ramírez et al. (2004) to detect the acoustic envelope, with a custom decision procedure to capture the different types of pauses of the speaker. The resulting voiced segements were then analyzed for both voice quality and pitch variation.

The resulting data were modeled under the assumption of

multimodality. A Support Vector Machine classifier (Cortes and Vapnik 1995) was used to measure the predictive value of the discovered patterns in the measurements of fluency, voice quality and pitch variation.

Fluency: VAD analysis

Fluency is measured as the number and duration of short and long unfilled pauses. To find pauses, the VAD algorithm of Ramírez et al. (2004) is used as follows.

Because the audio files in this study were recorded under various differing circumstances – some with a high, some with a very low Signal to Noise Ratio –, a per-file noise profile must be established prior to further computations. In this study, we used a scan for the best 500 ms noise window in the first 30 seconds of the audio. The noisiness of a window is equated to the distribution of the spectral energy in the spectrum bands of interest. The assumption is that noise contains less acoustic energy, and is distributed more randomly than speech.

In the initialization phase, the average in each band is taken of the 5 lowest scoring windows. To ensure robustness against audio artifacts, we discarded the extreme 0.001% audio samples as outliers in this and the next phases.

The long-term spectral envelope (LTSE) of each frame l is computed over a range of j samples, with the spectrum divided into k bands. The maximum value of the amplitude spectrum is recorded and then related to the noise profile built up during the initialization phase to establish the long-term spectral divergence (LTSD). We used overlapping windows of 13 frames, with a frame length of 10 ms.

In the decision phase, all LTSD-measurements are scaled and centered around their mean, computed over the 99% Highest Density Interval to exclude extreme outlier values. Let γ_l be the LTSD-value for frame l; the speech/non-speech decision for l is then made based on whether γ_l exceeds the noise profile computed in the initialization phase.

Speech quality: Cepstral Peak Prominence

This study uses Cepstral Peak Prominence (CPP, Hillenbrand and Houde 1996) to measure voice quality. A complex sound wave such as an utterance of human speech, is the sum of a number of sine waves with different amplitudes and frequencies. A Fourier transformation of the wave yields a spectrum of its constituting sine waves. This spectrum itself can be subjected again to an inverse Fourier transformation, yielding a *cepstrum*, a log spectrum of the log power spectrum of frequencies. The frequency of the sine waves that compose the spectrum ("*quefrencies*") represents the periodicity of spectral peaks.

If voice quality is affected, the signal will be less harmonic, which can be measured as reduced periodicity of the spectral peaks. Cepstral Peak Prominence, in this study computed on smoothed data and parametrized for connected speech (CPPS-s) measures the deviation (height) of the cepstrum that represents a speaker's fundamental frequency. CPPS has been shown to correlate with human judgements and with other measures of voicing instability (such as jitter and shimmer) in pathological speech (Fraile and Godino-Llorente 2014).

Pitch range: pitch trend deviations

Speech that is affected by dementia has been described as relatively monotone in terms of pitch variance. Speech and language therapists sometimes use the terms "robotic voice" and "monopitch", although there is not yet a satisfying way to quantify that perception. Pitch height is often measured as the fundamental frequency of a voice, expressed in Herz. In this study, the fundamental frequency is measured in *cents* instead. That measure is based on the relationship between tone differences and octaves: two tones are an octave (= twelve semitones, each divided in 100 $cents$) apart if their fundamental frequency doubles. Measurements in cents facilitate comparison of pitch height variations on a linear scale.

Audio recordings were sampled for pitch with a sampling rate of 5 milliseconds and a high-pass band filter to exclude measurements lower than 75 Hz.

Global and local trends were computed following the protocol as outlined in Matteson, Olness, and Caplow (2013). Outliers are identified using a modified z-score (Iglewicz and Hoaglin 1993). A windowed mean (local trend) is computed as the mean of measurements excluding outliers in a window of 5 seconds around a data point n. The local trend represent pitch changes that occur over the span of a few words. It is computed relative to a global trend, which represents the slow drift of pitch over a major part of the discourse. The prosodic variation that a speaker uses as linguistic device is measured as pitch deviations from the local trend. The range of deviations and its Probability Density Function is used to characterize the monotoneness in a discourse.

Results

VAD evaluation

In order to evaluate the algorithm, predictions on 40-second samples (n=10) were extracted from random positions in the audio files. The predictions were then compared to hand-labeled segmentations. The value of Cohen's kappa coefficient showed very good agreement (κ = 0.92).

The diagnostic ability of the VAD-implementation is evaluated under various choices of scaling factors in the decision step. For each of the samples, we computed which scaling factor yielded the highest F_1 measure (the harmonic mean of precision and recall); for this set of samples, the highest F_1 measure is when s is assigned a value in the range between 0.4-0.9.

The raw results are scaled, assuming their distribution is best approximated by the log-normal distribution (Campione and Véronis 2002), and binned. The resulting data can be described as the sum of multiple distinct Gaussians.

We estimate the parameters of each Gaussian using the `mixtools`-package (Benaglia et al. 2009) in R. Visual inspection of the histograms (plotted in Fig. 1), with their estimated Gaussians overlayed, shows that a mix of two Gaussians provides a good fit to the data.

The pause patterns for non-brain-damaged speakers, and those with AD, PD, SD are remarkably similar, with a short pause at about 100 ms and a long pause at around 350 ms. Speakers with the non-fluent kind of primary progressive

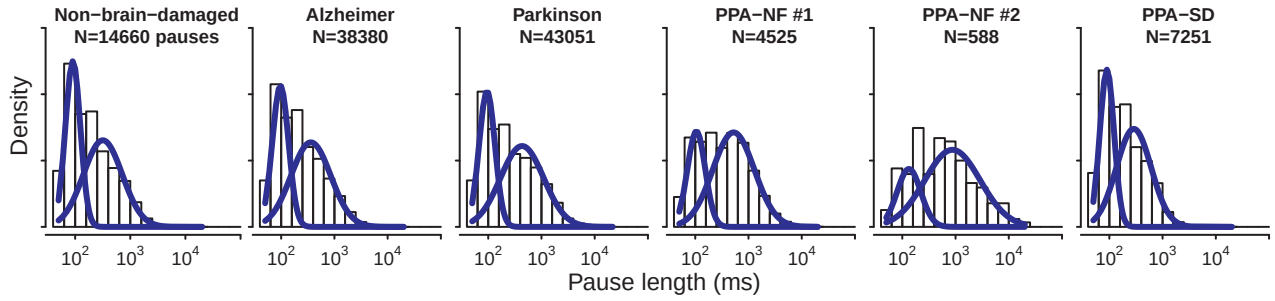


Figure 1: The frequency of pauses (as $\log_{10}$ ms) in all fragments, with a mixture of the estimated two Gaussians overlaid. Each bar represents the number of pauses with a given length, on a relative scale. Two non-fluent particiapnts (PPA-NF1 and PPA-NF2) are graphed separately.

aphasia lack the distinctive short pause peak, but have relatively more medium-to-long pauses.

The operational variables used in classification are the mean and SD of each of the estimated curves and their mix ratio.

Voice Quality

The descriptive statistics of the cepstral measure across the three topics and five groups shown that there were no significant differences between mean CPPS scores (where $p < .05$) across groups: $F(6, 34) = 0.86, p = 0.53$, or topics: $F(2, 68) = 2.26, p = 0.11$, or their interaction: $F(12, 68) = 1.69, p = 0.08$.

Pitch Trend Deviations

Groupwise differences of parameters between individuals with and without neurodegenerative diseases were evaluated through a side-by-side comparison using the Wilcoxon rank sum test. At 95%, the tests indicated that the range in segments of NBDs is not different from the range in any of the other participant groups. The relatively sample sizes decrease the statistical power of the test, but the results are in line with conclusions drawn from visual inspections of plots of pitch range of individual participants.

Machine Learning

The dataset with all variables (fluency, voice quality, prosody) was split in a training and test set (85%/15%). Three different one-vs-one SVM models were trained, with a linear, polynomial (degree 2) and RBF kernel. The latter yielded the highest accuracy, as expected with a small dataset, with $C = 8$ and $\gamma = 2$. The distribution of our labels is highly skewed due to the original setup of the study that produced the corpus. As evaluation metric, we use Area Under Curve, which is a relatively robust measure for skewed data. AUROC curves of the classifiers predicting *NBD* vs. *AD*, *NBD* vs. *neurodegenerative disease* and *NBD* vs. *PPA-NF* are shown in Fig. 2. All other classifiers had accuracies lower than their No Information Rate, indicating that the classifier performed at less than chance.

Discussion

In this study, prosody measurements were computed in conversational speech in an automated way. The results show that a classifier trained on these measures can detect NBD vs. AD (with marginal confidence), NBD vs. neurodegenerative diseasee and NBD vs. PPA-NF (with high confidence). Differences between or within the other categories were not detectable using automated measurements. The good performance of the PPA-NF classifier suggests that the speech-pause pattern in speech of individuals with PPA-NF is sufficiently different from that of individuals from the other classes to serve as a good predictor.

A relatively simple VAD-algorithm forms the basis of the non-fluency detection. Clinical characteristics of non-fluency typically include both filled and unfilled pauses. However, the algorithm in this study is only sensitive to unfilled pauses. This suggests that the occurrence of unfilled pauses alone is enough to detect this condition. Followup experiments may investigate the correlation between unfilled and filled pauses, and whether one may be used to predict the other.

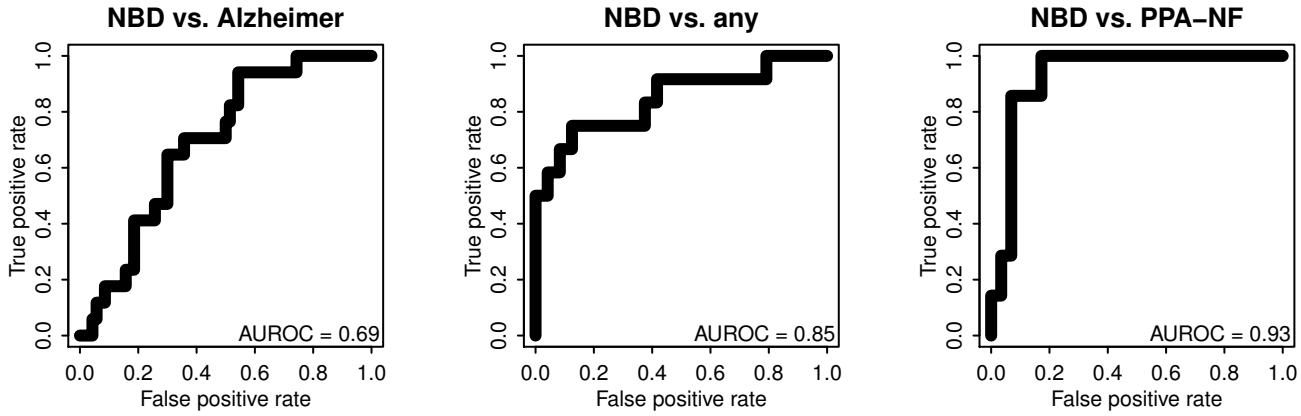
A limitation of this study is the small number of participants. Small samples decrease the power of statistical tests and increase the probability of type I or type II errors. The individual variables show a large variance. A larger sample size is required to make more rigorous claims about the performance of the classifier.

The classifier could not distinguish speech from dementia types associated with decreased pitch range or voice quality. Post hoc testing showed that none of the measures shows differences significant enough to reject the null hypothesis of them being drawn from the same distribution.

Conclusions and Future Work

Some prosodic features can be be used in a classifier of certain forms of dementia. A classifier trained on the output of a basic VAD algorithm can distinguish non-fluent PPA participants from other participants. Accuracy of the classifier is significantly larger than that of a "no information" strategy. PD and PPA-SD participants were not distinguishable from controls by the classifier.

The measurements of pitch range and voice quality can



(a) (b) (c)

Figure 2: (a) ROC curves for the performance of the SVM classifier for different binary prediction tasks, after cross validation.

converge more when the sample size is increased, perhaps yielding more informative features for a classifier to detect the AD and PD participants. Prosodic measurements are easy to obtain, and relatively robust, but they prove limited in their use for this domain.

The addition of lexical and (morpho-)syntactic information will most likely improve the classifier. The approach in this study serves as a pilot for a study that will include more participants. The use of automated measurements can eventually lead to software instruments that can be used in clinical practice for screening and diagnosis.

Acknowledgments

This work is part of the research programme "Doctoral Grant for Teachers - 2015 BOO" with project number 023.0070.61, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). Roelien Bastiaanse is partially supported by the Center for Language and Brain NRU Higher School of Economics, RF Government grant, ag. Nº 14.641.31.0004.

References

Baken, R. J., and Orlikoff, R. F. 2000. *Clinical measurement of speech and voice*. Cengage Learning.

Benaglia, T.; Chauveau, D.; Hunter, D. R.; and Young, D. 2009. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software* 32(6):1–29.

Campione, E., and Véronis, J. 2002. A large-scale multilingual study of silent pause duration. In *Speech prosody 2002, international conference*.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine learning* 20(3):273–297.

Forrest, K.; Weismer, G.; and Turner, G. S. 1989. Kinematic, acoustic, and perceptual analyses of connected speech produced by parkinsonian and normal geriatric adults. *The Journal of the Acoustical Society of America* 85(6):2608–2622.

Fraile, R., and Godino-Llorente, J. I. 2014. Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control* 14:42 – 54.

Gorno-Tempini, M. L.; Dronkers, N. F.; Rankin, K. P.; Ogar, J. M.; Phengrasamy, L.; Rosen, H. J.; Johnson, J. K.; Weiner, M. W.; and Miller, B. L. 2004. Cognition and anatomy in three variants of primary progressive aphasia. *Annals of neurology* 55(3):335–46.

Hillenbrand, J., and Houde, R. A. 1996. Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. *J Speech Hear Res* 39(2):311–21.

Iglewicz, B., and Hoaglin, D. C. 1993. *How to detect and handle outliers*, volume v. 16. Milwaukee, Wis.: ASQC Quality Press.

Kato, S.; Endo, H.; Homma, A.; Sakuma, T.; and Watanabe, K. 2013. Early detection of cognitive impairment in the elderly based on bayesian mining using speech prosody and cerebral blood flow activation. *Conf Proc IEEE Eng Med Biol Soc* 2013:5813–6.

Matteson, S. E.; Olness, G. S.; and Caplow, N. J. 2013. Toward a quantitative account of pitch distribution in spontaneous narrative: method and validation. *J Acoust Soc Am* 133(5):2953–71.

Ramírez, J.; Segura, J. C.; Benıtez, C.; De La Torre, A.; and Rubio, A. 2004. Efficient voice activity detection algorithms using long-term speech information. *Speech communication* 42(3):271–287.

Roark, B.; Mitchell, M.; Hosom, J. P.; Hollingshead, K.; and Kaye, J. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech and Language Processing* 19(7):2081–2090.