Detecting Intoxication in Speech

Matthew Marge

SLTC Newsletter, October 2011

http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2011-10/detecting-intoxication-in-speech/

Researchers at Columbia are investigating ways to automatically detect intoxication in speech. William Yang Wang, currently a PhD student at Carnegie Mellon that worked on this team while a Master's student, discussed the project and its goals with us.

OVERVIEW

Imagine a world where DUI's (driving under the influence violations) never occurred. How can this happen? Traditionally devices like breathalyzers can detect intoxication, but these tools are expensive and impractical for "passive operation" in vehicles. One arguably more practical alternative is that the cars themselves listen to the driver, detect that the potential driver is intoxicated, and prevent the car from starting. Researchers at Columbia are investigating a crucial problem in this space - ways to automatically detect intoxication in speech.

In a recent presentation at the Intoxication Sub-Challenge at Interspeech, the Columbia groupdiscussed their approach [1, 2]. Their key intuition was to treat the intoxication of a person's speech similarly to a person's accent, and apply existing automatic methods to accent detection. The system they built could detect differences in the phonetic structure of sober and intoxicated speech.

William Yang Wang, currently a PhD student at Carnegie Mellon that worked on this team while a Master's student, discussed the project and its goals with us.

SLTC: What kind of system did your group build?

Wang: At Columbia, we have studied speaker states that do not map directly to the classic or even derived emotions: charismatic speech, deceptive speech, depression, and levels of interest. In this work, we are interested in building a system that can automatically classify intoxicated speech from sober speech, given only a very short sample from the speaker. This is a crucial task from the point of views of public safety and social welfare. In the United States, there were 164,755 alcohol related fatalities in a past window of ten years (1999-2008). A system to detect a person's level of intoxication via minimally invasive means would not only be able to significantly aid in the enforcement of drunk driving laws, but also ultimately save lives. The system should be easily applicable to the domain of in-car intoxication detection, as well as automatic speech recognizers and spoken dialog systems.

How does it work?

The traditional approaches to speaker state detection include two steps: the extraction of low-level acoustic features (e.g. MFCC), and n-way direct classification or regression using maximum margin classifiers. However, this might not always yield optimum results because the model is too simple, and might not be able to capture subtle phonetic differences between drunk and sober speakers. When I first heard of this shared task from my advisor Julia Hirschberg, I was thinking that maybe we could cast this problem as a problem that we are more familiar with: an accent identification problem, where we assume intoxicated speakers speak a slightly different accent of the same language than sober speakers. In fact, Fadi Biadsy has just finished his PhD thesis on automatic dialect and accent identification at Columbia, so we thought it would be an interesting idea to run his system on this dataset.

Which properties of speech did you study in your work? Which yielded the best results?

We have investigated high level prosodic, phone duration, phonotactic, and phonetic cues to intoxicated speech. The first thing we looked at was the n-gram frequencies of prosodic events in an utterance, where we hypothesize energetic and depressed intoxicated speakers might use higher and lower rates of emphasis (pitch accents) than sober speakers. In addition to this, a greater rate of disfluencies or intonational phrase boundaries might occur in intoxicated speech. Unfortunately, due to data sparsity and other problems, the final results using the prosodic approach were not as good as what we had in previous emotional speech classification tasks. The second property we looked at was the changes of phone durations in intoxication state. In our study, a simple vector space model of phone durations has been shown to be very effective, and the result was even better than the acoustic baseline that used more than 4000 low-level acoustic features. We then study a phonotactic based language/accent identification technique for this task, where we believe intoxicated speakers might use certain phones or words more frequently than others. The experiment results showed that our hypothesis was correct. Finally, we have investigated a state-of-the-art accent identification technique where we hypothesize intoxicated speakers realize certain phones differently than sober speakers. We built adapted Gaussian Mixture Models - Universal Background Models (GMM-UBMs) for each phone type in the data, generated GMM supervectors and used an upper-bound KL-divergence based linear kernel SVM to compare the similarity between two utterances. Our final system that combines this method with the phonotactic method achieved the best result, and it was significantly better than the majority and official baseline of the sub-challenge.

What kind of impact do you foresee this work having?

The most important impact in this work is that we successfully cast the unknown problem of intoxication detection to the problem of language and accent identification, which we have known for years. In the field of language and accent identification, NIST organizes language recognition evaluation workshops every two years, and some of the best systems from the workshop have already reached 3-4% equal error rate in a 30-second close-set dialect/accent identification task. This is definitely encouraging news for the emotional speech community, because the state-of-the-art emotion and speaker state detection systems still have much lower accuracy and cannot be compared with mature speech techniques like speech recognition, speaker and language identification. We also believe that it is necessary to investigate more prosodic cues to the speaker state and emotion detection tasks.

Can you tell us a bit about the Intoxication Sub-Challenge at Interspeech this year? What were the goals of the challenge? What were your personal impressions of the sub-challenge?

Definitely. Interspeech has organized three shared tasks on emotion, paralinguistic, and speaker state detection in the past three years respectively. All of them were very interesting tasks. This year, the organizers thought they would like to focus on a rarely studied speaker state: intoxication, and the goal was to identify interesting features that could characterize intoxication and approaches that can separate intoxicated speech efficiently. Personally speaking, I thought the organizers did a relatively good job in organizing the event, and it was a great idea to work on a shared dataset with researchers around the world. The challenge results are not the most important thing, but this kind of competition can definitely put our field forward. In terms of the suggestions, we might consider the following two questions in the future challenges: (1) what are the most appropriate evaluation metrics? (2) how could we preserve the authenticity of the data when creating training, development and test sets for the challenge?

We look forward to hearing more about intoxication detection in speech in future challenges!

REFERENCES

- [1] Bjorn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, Jarek Krajewski. "The INTERSPEECH 2011 Speaker State Challenge", in Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), Florence, Italy 28-31 Aug., 2011.
- [2] Fadi Biadsy, William Yang Wang, Andrew Rosenberg, Julia Hirschberg, "Intoxication Detection using Phonetic, Phonotactic and Prosodic Cues", in Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), Florence, Italy 28-31 Aug., 2011.

If you have comments, corrections, or additions to this article, please contact the author: Matthew Marge, mrma...@cs.cmu.edu.

Matthew Marge is a doctoral student in the Language Technologies Institute at Carnegie Mellon University. His interests are spoken dialog systems, human-robot interaction, and crowdsourcing for natural language research.

Editor's note: For more information, readers are referred to the papers presented at the 2011 Interspeech Speaker State Challenge: Session 1 and Session 2. The challenge is documented in more detail here. For the intoxication challenge, it is interesting to note the wide variety of approaches of taken. For example, the winner of the Intoxication sub-challenge prize -- Bone et al, Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors -- relied on hierarchical organization of speech signal *features, the use of novel speaker normalization techniques as well as the fusion of multiple classifier subsystems. The adopted features included spectral cues long used for speech recognition, and signal features of prosody, rhythm, intonation and pitch, and also voice quality cues such as hoarseness, creakiness, breathiness, nasality, and quiver, all inspired by phonetics research and speech analysis results in the published literature. -- Jason Williams, 2010-10-28.*