Using Prosodic Features to Prioritize Voice Messages

Tim Polzehl
Deutsche Telekom Laboratories
Technische Universität Berlin
Berlin; Germany
tim.polzehl@telekom.de

Florian Metze
Deutsche Telekom Laboratories
Technische Universität Berlin
Berlin; Germany
florian.metze@telekom.de

ABSTRACT

This paper presents a preliminary study to classify voice messages left by callers to a call center into semantic categories by evaluating prosodic information on about 10 seconds of speech only. We analyze the variation of suprasegmental features such as F_0 and loudness of more then 400 calls and find that this information should be sufficient to map calls to semantic categories defined previously, at least for the purpose of prioritizing further call processing. Prioritization is important in order to make efficient use of resources that would lay idle otherwise.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Voice I/O*

General Terms

Content analysis based on meta data

Keywords

speech processing, speaker characteristics, semantic categorization

1. INTRODUCTION

There is currently a trend to employ speech-to-text (STT) technology in large-scale, commercial call centers for purposes such as call mining and speech analytics. In these applications, a transcription of the speech data is generated not using a grammar, but using a statistical language model (SLM). The output is used to generate statistics from the sheer volume of calls and derive answers to questions such as *"what are our customers concerns"*, *"what do our customers think about our service"*, or to monitor the quality of the customer experience in off-line mode.

Speech search workshop at SIGIR July 2008; Singapore

STT based on SLMs is significantly more expensive computationally than standard grammar-based speech recognition components of interactive voice response systems. To make efficient use of available resources, i.e. CPU cycles in call centers, calls are usually processed off-peak, i.e. at times when machines would otherwise be idle anyway.

While speech analytics is not generally geared towards deriving information about an individual call, a company might want to react proactively to information discovered in a specific call. In order to maximize the usefulness and to avoid the customer calling again, a quick turnaround time is needed. Given the CPU constraints, it is desirable to classify calls using non-verbal information, which can be computed cheaply. Then, calls could be prioritized accordingly for further processing. Also, we expect this off-line information to be useful for designing interaction patterns for future adaptive on-line systems [2, 4].

2. TEST DATA

Our experiments were conducted on an internal set of 439 utterances, collected from live customers after they had interacted with the all-purpose production customer care call routing platform, and an agent. During an automatic after-call survey, these customers had given the worst possible mark to the service provider in general using a 5-grade scale, and agreed to describe the reason for their rating in a "voice message" to the provider. Messages are in German and usually contain 7 to 10 seconds of telephony quality speech.

These calls were transcribed manually and tagged as belonging to one or more of 9 semantic categories. Categories were defined on this set of data with the intent of being able to treat calls belonging to different categories individually. In this paper, we restrict ourselves to the following 5 categories, because they appear frequently and we hope to exploit them for improved scheduling in the future:

Cat 1 Customer says he did not have a proper conversation with an agent and therefore cannot rate it (15.7%)

Cat 2 C. says automatic interview began before interaction with agent was terminated properly (18.5%)

Cat 3 C. describes the original reason for the call (18.2%)

Cat 4 C. expresses unhappiness with the company, customer care, or that his problem had not been dealt with satisfactorily (31.9%)

Cat 5 C. says he does not want or is unable to deal with a speech dialog system (7.9%)

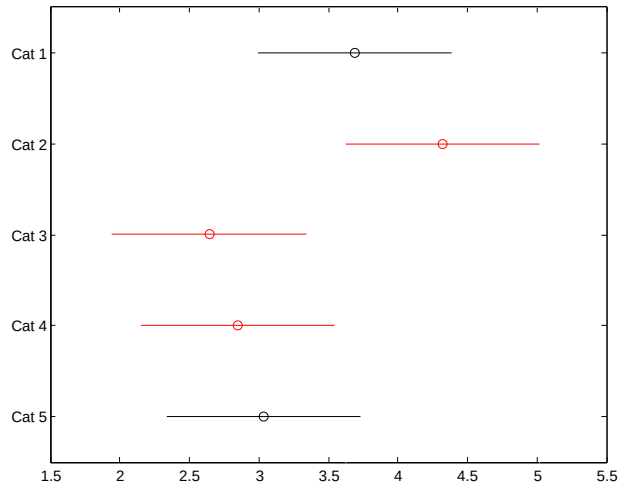

Figure 1: Standard deviation of F_0 including confidence intervals for our 5 most important semantic classes. Red color denotes categories different from at least one other category with $p < 0.05$.

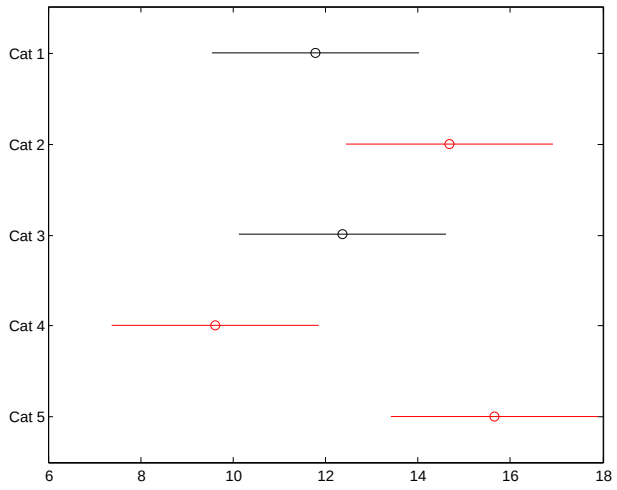

Figure 2: Standard deviation and associated confidence intervals of loudness for the 5 most important semantic classes found in our data-set.

3. EXPERIMENTS

Speech signals as recorded by the voice platform were processed with Praat (`http://www.praat.org`) using 10ms frame shift to derive prosodic information.

Figure 1 shows the means and distribution of confidence intervals for the standard deviation of pitch (F_0) for calls belonging to the 5 semantic categories described above. Pitch was extracted using Boersma's algorithm [1], which searches over a multitude of F_0 candidates and uses the Viterbi algorithm to find the best path. In order to optimize performance on narrow band telephony data, we adjusted silence and voice-ratio thresholds (costs of octave error). Only voiced frames were considered for calculation of the standard deviation. Given the quality of telephone speech, we are working on improved methods to detect and track F_0 and expect even better results in the future.

Cat 1 and Cat 2 contain utterances of callers who are surprised to be confronted with the after-call interview, so there is a high activation of the "variation of F_0" feature. Quality management should follow up on these issues immediately.

Figure 2 shows our results for perceptual loudness, measured using the algorithm proposed and implemented in the prosogram [3]. Similar to common spectral analysis, the speech signal is being converted to the frequency domain. The difference at this point is that an excitation parameter is calculated on the basis of a cochleagram, which represents the excitation pattern of the basilar membrane and is calculated for every time frame and every Bark frequency.

We find significant differences between loudness values of our semantic categories Cat 4 and Cat 2/ Cat 5. Here, it seems that the primary concern of the caller is the current confrontation with an unwanted automatic system, which manifests itself in a high variation of loudness. These issues should only be followed up manually.

We are currently investigating further features and characteristics and will also work on spotting keywords as well as examining non-lexical speech parts. Moreover, the scope of information sources will eventually be extended by further acoustic and dialog dependent characteristics.

4. CONCLUSIONS

This paper presented initial results of prosodic analysis of speech recorded in an automatic survey after callers had spoken to an agent in a call center. We found significant differences in variations in F_0 and loudness between calls clustered into different semantic categories, i.e. customers expressing a general concern (i.e. Cat 4) and customers which had been treated improperly just briefly ago (Cat 2). We intend to further study these effects and compare them with insights gained from interactive systems. This approach should allow us to prioritize and schedule automatic transcription of these utterances, which is still computationally expensive. It enables targeted load balancing and gives opportunity to conduct immediate follow-ups with customers. We are also starting to work on automatic classification of messages.

5. ACKNOWLEDGMENTS

The authors would like to thank Caroline Clemens for transcribing and annotating the data.

6. REFERENCES

[1] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, 17:97–110, 1993. Amsterdam; The Netherlands.

[2] F. Burkhardt, F. Metze, and J. Stegmann. *Advances in Digital Speech Transmission*, chapter Speaker Classification for Next Generation Voice Dialog Systems. Wiley, Jan. 2008.

[3] P. Mertens. The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In B. Bel and I. Marlien, editors, *Proceedings of Speech Prosody 2004*, Nara; Japan, Mar. 2004. ISCA.

[4] F. Metze, R. Englert, U. Bub, F. Burkhardt, and J. Stegmann. Getting closer – tailored human-computer speech dialog. *Universal Access in the Information Society*, 2008. Springer, Heidelberg. To appear.