Detecting Urgency in Speech with Personalised Acoustic Features

Jakob Landesberger
Mercedes-Benz AG & University Ulm
jakob.landesberger@daimler.com

Ute Ehrlich
Mercedes-Benz AG
ute.ehrlich@daimler.com

Abstract

Finding an appropriate reaction to a spoken utterance with multiple intents is not easy for a spoken dialogue system. One way to simplify the dialogue is to prioritize urgent intents. For this purpose, the urgent part of an utterance must be reliably identified. This is possible purely with acoustic features directly from the audio signal. In this work, we consider the acoustic features for each speaker individually and show that the recognition can be improved by this personalization.

1 Introduction

Utterances in dialogue serve often more than one communicative function. Like giving feedback about the understanding of a question and answering the question in a single utterance. The ability of humans to easily process such multiple communicative functions and to react accordingly, allows for a swift and effective communication (Lemon et al., 2002). If an utterance contains two sequentially occurring intents, each relating to a different task or activity, it is often referred to as multi-intent (MI). Such utterances also occur in interactions with a machine (Shi et al., 2019). If there is a need for further clarification of all mentioned intents, it can be difficult for a system to find a suitable answer. Answering with a MI, too, can produce long utterances, which can be cognitively very demanding (Landesberger and Ehrlich, 2019a). To solve this problem, we proposed a model for prioritizing tasks when dealing with MI utterances (Landesberger and Ehrlich, 2019b). The task, referred by an intent, is prioritized according to various criteria and the remaining tasks are postponed. After solving the prioritized task, the postponed tasks can be resumed. One criterion for prioritisation is urgency. A task is urgent if the task has to be completed in a short amount of time, because if not, it loses relevance or other negative consequences occur. *"What do I do with the zucchini? Oh, the pan is hot. What is coming in now?"* Certainly, the speaker would be frustrated if the first mentioned task is considered before the second one. In order to realise a suitable dialogue behaviour for this problem, it must first be detected which part of the utterance is urgent.

2 Urgency Corpus

In order to develop a method for the detection of urgency in utterances we use data obtained with the help of the game "What is it?" (Landesberger et al., 2020). The players' task was to find a special symbol on a matrix. To exclude wrong symbols, questions that could be answered with yes or no had to be asked to a simulated system. This task was regularly interrupted by an easier but time-limited, urgent task. This resulted in a Corpus of spoken urgent and non-urgent utterances. Since the task types alternated quickly, many utterances contained both urgent and non-urgent intents, just like in the example in section 1. Although, the different utterances are not very comparable regarding the content, we decided on a semantically independent approach, to later tackle real world problems. Accordingly we tried to detect urgency based on acoustic features from the audio signal.

3 Acoustic Features

In our previous work (Landesberger et al., 2020) the analysis of 108 acoustic features like pitch, intensity or MFCCs, showed that depending on the speaker's phase it was more or less difficult to identify urgency. If the speaker switches from a non-urgent intent to an urgent intent, he is in the phase Transition (TRA). Conversely, if he changes from an urgent intent to a non-urgent one, he is in the Decline (DEC) phase. For each of these phases and a general distinction (ALL) in which all utterances were analysed, we first identified a feature subset that contains as few irrelevant and redundant (John et al., 1994) features as possible. For this purpose we used three different estimators with the recursive feature elimination method (Granitto et al., 2006): Logisitc Regression Estimator (LRE (Hosmer Jr et al., 2013)), Random Forest Estimator (RFE (Breiman, 2001)) and Gradient Boosting

Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue, July 18-19, 2020, Online, hosted from Massachusetts, USA.

		RFC	GBC	LRC	KNNC	MLPC	SVMC
ALL	LRE	.883 +1.20%	.795 +4.26%	.774 +7.27%	.814 +1.11%	.798 +10.28%	.888 +13.66%
	RFE	.883 +1.85%	.795 +5.12%	.774 +8.82%	.814 +1.71%	.801 +12.27%	.888 +14.00%
	GBE	.883 +1.93%	.796 +4.35%	.775 +8.20%	.814 +1.06%	.799 +11.67%	.887 +14.61%
TRA	LRE	.779 +4.35%	.769 +2.17%	.782 +8.32%	.712 +5.22%	.784 +13.70%	.767 +1.02%
	RFE	.774 +2.55%	.770 +2.44%	.782 +8.61%	.712 +6.63%	.787 +16.15%	.767 +2.58%
	GBE	.787 +4.40%	.779 +3.67%	.781 +8.42%	.687 +3.44%	.777 +8.95%	.784 +3.84%
DEC	LRE	.883 -1.38%	.891 -0.22%	.906 +3.33%	.878 +1.80%	.895 +1.25%	.891 -0.43%
	RFE	.889 -1.08%	.889 -1.08%	.895 +2.69%	.886 +2.95%	.886 +0.98%	.891 -1.63%
	GBE	.887 -0.92%	.890 -1.47%	.897 +1.97%	.877 +3.20%	.893 +1.22%	.901 +0.07%

Table 1: Accuracy and personalisation improvement in percent to detect urgent user requests over all the data (*ALL*), during the switch from non-urgent to urgent utterances (*TRA*) and during the switch from urgent to non-urgent utterances (*DEC*) for each classifier estimator combination

Estimator (GBE (Friedman, 2001)).

In order to automatically distinguish between urgent and non-urgent intents, we trained and tested six different classifiers with different supervised learning algorithms: Random Forest Classifier (RFC (Breiman, 2001)), Gradiant Boosting Classifier (GBC (Friedman, 2001)), Logistic Regression Classifier (LRC (Hosmer Jr et al., 2013)), K Nearest Neighbours Classifier (KNNC (Peterson, 2009), Multi-Layer Perceptron Classifier (MLPC (Pal and Mitra, 1992)) und Support Vector Machine Classifier (SVMC (Vapnik, 2013)).

We analysed the accuracies achieved by each estimator and classifier combination distinguishing between urgent and non-urgent intents. Each of these values was validated by a 10-fold stratified cross validation method. For the Transition phase, the SVMC in combination with LRE showed the best result with an accuracy of .759. During Decline, the best result was achieved with the SVMC and the RFE with an accuracy of .909. The LRE in combination with the RFC performed best with an accuracy value of .873 in the general distinction.

4 Personalisation

Acoustic features in speech, change strongly depending on the speaker. For example, men usually speak with a lower pitch than female speakers do. Accordingly, the absolute or normalized values of certain features are not best suited for the detection of urgency. If a male speaker speaks higher than he would normally speak during an urgency, this value can still be below the normal pitch of a female speaker. Accordingly, the consideration of individual differences and corresponding relative, personalised variations in acoustic features could improve the recognition of urgency. In the examined data set 40 participants uttered 10737 intents. With on average of 268.4 per test person, we adopted the min-max normalization formula to include the individual variance for each participant p and feature f:

$$x'_{pf} = |\frac{x_{pf} - min(x_{pf})}{max(x_{pf}) - min(x_{pf})}|$$

The new values x'_{pf} rank on a scale from 0 to 1. 0 corresponds to the lowest measured feature value of one participant in all his utterances and 1 to the highest. Based on the recalculated values, the already evaluated classifiers and estimators were retrained and re-evaluated. The results are shown in Table 1. In addition to the accuracy of the individual classifier estimator combinations, the percentage change compared to the non-personalized data is shown. It is noticeable that there is no improvement if the accuracy was already high before. During the Phase Decline (DEC) even minimal deteriorations occur. In the other comparisons (ALL and TRA) all values improved. Especially the SVMC and the MLPC benefited from the personalization.

5 Conclusion

In this paper we show that adaptation to the speaker can help to identify urgent intents in utterances more reliably. For this adaptation, however, the speaker must already be known to the system. Adaptation to more general characteristics, such as gender, might be practical, too.

To further improve the detection of urgency, we plan to consider other aspects of the spoken utterance and the context of the utterance, too. We assume that certain words such as deictic expressions, the task addressed, and the current situation of the speaker can all provide indications for urgency.

References

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Pablo M Granitto, Cesare Furlanello, Franco Biasioli, and Flavia Gasperi. 2006. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2):83–90.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*, volume 398. John Wiley & Sons.

George H John, Ron Kohavi, and Karl Pfleger. 1994. Irrelevant features and the subset selection problem. In *Machine Learning Proceedings 1994*, pages 121–129. Elsevier.

Jakob Landesberger and Ute Ehrlich. 2019a. Finding a meta-dialogue strategy for multi-intent spoken dialogue systems. In *Proceedings of the 3rd International Conference on Computer-Human Interaction Research and Applications - Volume 1: CHIRA,*, pages 171–176. INSTICC, SciTePress.

Jakob Landesberger and Ute Ehrlich. 2019b. Towards finding appropriate responses to multi-intents - spm: Sequential prioritisation model. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue*, London, United Kingdom. SEMDIAL.

Jakob Landesberger, Ute Ehrlich, and Wolfgang Minker. 2020. Do the urgent things first! detecting urgency in spoken utterances based on acoustic features. In *Proceedings of the 28th Conference on User Modeling, Adaptation and Personalization*, UMAP '20. Association for Computing Machinery.

Oliver Lemon, Alexander Gruenstein, Alexis Battle, and Stanley Peters. 2002. Multi-tasking and collaborative activities in dialogue systems. In *Proceedings of the 3rd SIGdial workshop on Discourse and dialogue-Volume 2*, pages 113–124. Association for Computational Linguistics.

Sankar K Pal and Sushmita Mitra. 1992. Multilayer perceptron, fuzzy sets, classifiaction.

Leif E Peterson. 2009. K-nearest neighbor. *Scholarpedia*, 4(2):1883.

Chen Shi, Qi Chen, Lei Sha, Hui Xue, Sujian Li, Lintao Zhang, and Houfeng Wang. 2019. We know what you will ask: A dialogue system for multi-intent switch and prediction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 93–104. Springer.

Vladimir Vapnik. 2013. *The nature of statistical learning theory*. Springer science & business media.