Application of Speaker Classification in Human Machine Dialog Systems

Felix Burkhardt[1], Richard Huber[2], and Anton Batliner[3]

[1] T-Systems International, Goslarer Ufer 35, 10589 Berlin, Germany,
felix.burkhardt@t-systems.com,
http://www.t-systems.com
[2] Sympalog Voice Solutions GmbH, Karl-Zucker-Str. 10, 91052 Erlangen, Germany
huber@sympalog.de,
http://www.sympalog.de
[3] Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Martensstr. 3, 91058 Erlangen, Germany,
batliner@informatik.uni-erlangen.de,
http://www.informatik.uni-erlangen.de

Abstract. This chapter deals with the application of automatic speaker classification in human machine dialog systems based on telephone operation. In a first step we introduce a taxonomy based on three features that such systems might have. We explain the features, namely *online*, *mirroring*, *critical* and their respective counterparts get explicated and are than used to characterize a part of the exemplary applications that illustrate the benefit of that approach. Furthermore prototypical application scenarios are described that shall illustrate the vast possibilities to utilize automated speaker classification in dialog applications.

Key words: speaker classification, applications, voice portals, taxonomy

Contents and Motivation

In human-human communication speaker classification takes place at all times and is very valuable for the communication process, as people constantly adapt their manner of speaking based on the assessment they have of their counterpart's age, gender, mood or mother tongue. Incorporating such strategies into (semi-)automated voice services can be very helpful to extend their benefit. On the other hand, the possibility of automated speaker classification based on telephone-calls makes new applications possible based on this feature in itself.

This chapter shall envisage some application that utilize speaker classification in a telecommunication scenario. We will propose a taxonomy of such applications based on features like online/offline or mirroring/non-mirroring. Further we will introduce a set of application scenarios and discuss in parts their place in the taxonomy. Further ideas for applications shall illustrate the many possibilities to utilize speaker classification with automated dialog systems.

A Taxonomy for Speaker Classification Applications

In [1] we introduced a taxonomy to distinguish between applications that utilize emotional awareness. Because emotion recognition can be seen simply as a subset of speaker classification, this approach is extensible with respect to other speaker classifications based on age, gender or accent.

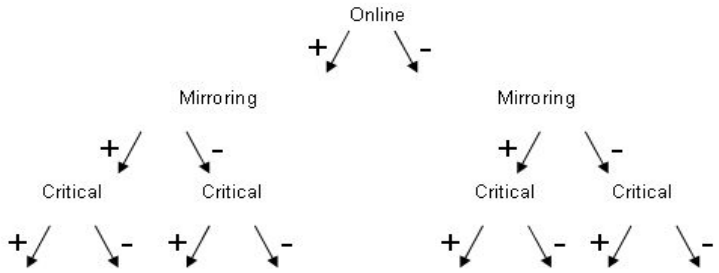

Fig. 1. A possible taxonomy of speaker classification applications.

Such a taxonomy can be very useful to think up possible applications that utilize speaker classification by playing around with the inclusion or exclusion of certain features. In figure 1 a possible taxonomy is depicted as a tree. Now if we have an application that resides in one leaf of that tree, moving it to another place might give us ideas to think up a different application. Examples are given below. In the following paragraphs we introduce three binary features that could be used for such a taxonomy.

Online vs. Offline: With an online system the classification is made immediately while interacting with user, while with an offline system the classification takes place delayed, after the actual interaction, based on logged audio data. Take as an example an ad sponsored telephone service that presents target group optimized advertising based on user classification. In an online version the system would decide with each call which target specific advertisement to display whereas an offline version of the same application would simply analyze the main user groups, perhaps analyzed for certain time-intervals or branches of the dialog, and present user-targeted advertising based on that decision.

Mirroring vs. Not Mirroring: This distinction is based on whether the user gets a direct feedback based on the classification or whether he/she is not directly aware of it. Take as an example a system that analyzes the perception of accent. In a mirroring version such a system could help a language student to improve the pronunciation[4] while a non-mirroring variant might be used to offer the caller a dialog in his/her mother tongue.

[4] Although we admit that it is not straightforward to imagine this as a service offered via telephone.

Critical / Not Critical: If we talk about a *critical* application we mean that the feature "speaker classification" is indispensable for the success of the application's purpose. Of course all applications whose main feature consists of the speaker classification, like the example of the language-student that gets supported by accent detection, fall into that category.

A further differentiation of applications can be based on the speaker features that are classified. An application that classifies callers with respect to their age might make sense also with other features like accent, gender or emotion. E.g. the obvious idea to offer callers an optimized dialog design according to their age group can be mapped to emotion recognition in the anger-detecting voice portal scenario.

Application scenarios

In the following we will introduce a set of applications that utilize speaker classification. Some of them were repeatedly envisaged in scientific literature, others have been mentioned already in the public media and few are even deployed as real-world applications. Many applications based on emotion recognition are further discussed in [2] or [1]. As such they can be taken as prototypical applications that stand for a family of related ideas. By applying the taxonomy mentioned in Section 1 the set can be enlarged by thinking up applications that differ with respect to a certain feature or utilize the classification based on a different speaker feature.

Dispatching callers to trained agents

This describes in a generic way the idea to classify the customers of a call center and forward them to an agent whose profile matches the caller's class. One obvious example would be the language of the caller in a multilingual call-center, e.g. a credit-card hotline. Another famous application foresees the measurement of anger in the caller's voice (e.g. for a complaint hotline), so that very angry customers can get handled by specially trained agents [3].

But generally spoken, the idea to let a specially talented agent take care of certain speaker classes can be generalized to classes like men, women, elderly, kids, strangers, or people coming from a specific part of country.

Adapted Dialog Design

If we take automated voice portals into account, dialog design becomes an important issue. State of the art technology in language understanding and artificial intelligence does not yet allow for totally free and open dialogs in human computer interaction (HCI). Dialog design comprises the way how the dialog flow is designed, i.e. which grammars are activated, which prompts will be played, which choices can be made by the user and which feedback strategies are implemented.

In the case of static speaker classification like age or gender one possible application in this context would be to implement several designs and activate the one that fits best to the current user profile. This might consist of very subtle changes, for example elderly customers might prefer a slower speech rate in the system's prompts. A misclassification would then not lead to a perceptible difficulty for callers, resulting in a non-critical application with respect to the above mentioned taxonomy.

On the other hand a dynamic speaker classification like e.g. emotion or language could be used to adapt the dialog dynamically to a change in the user's state. One of the most famous examples for such an application consists of the emotion-aware voice portal that detects user's anger and tries to soothe him/her by comforting dialog strategies as described in [3]. With this example a misclassification might lead to serious problems as callers that were not angry will probably get angry if accused unjustly. In that sense the application can be seen as critical.

Target-group specific advertising

Analogous to the development of Internet services it is foreseeable that a rising number of telephone services will by financed by advertising. Knowledge of the user group can help a great deal with respect to product choice and way of marketing.

According to the taxonomy this could be used as an online application if the users get classified in the beginning of their conversation and tailored advertisement is presented to them in the course of further interaction, e.g. while waiting for a connection.

As an offline application on the other hand the main user groups of a specific voice-portal or branch of a voice portal could be identified during a data collecting phase and then later advertisement targeted for the major user group be chosen.

This is not a critical application as the main target of the interaction (a user gaining information) would not be endangered by the choice of an inappropriate advertisement.

It is also not mirroring as the user is not directly aware of being classified.

Adapted Persona Design

With "persona design" we describe an automated voice-portal's character as a virtual persona. The character is represented by the wording of the prompts, the sound of the systems voice[5] and the expressions the grammar consist of. The design of such a "persona" can be enhanced to a great deal by background music or other sound-effects. Encountering an inconsistent persona design can

[5] Irrespective of the fact whether it's playback of prerecorded prompts or speech synthesis.

be very confusing in an interaction, just like speaking to a person with multiple personalities.

Usually the persona design is influenced by the topic the voice-portal application is all about, e.g. a banking application will prefer a serious persona, perhaps an elderly gentleman whereas a ring tone download application might use a trendy younger girl character. But speaker classification makes it possible to design one and the same application with different personas and activate them based on a prior application.

Analogous to the target-group specific advertising application this would be conceivable in an online version and an offline version, where the major user groups are identified in a preliminary data collection phase.

Market analysis of target groups

Even if the system's reactions are not influenced by the speaker classification, a knowledge of which group called when or was interested in which product can be invaluable for marketing strategists. Because it's not important to gain such a knowledge during the course of interaction, this could be realized as an offline application, thereby allowing the use of more sophisticated classification algorithms that don't need to follow real time requirements.

Call Center Quality Management

Call center managers have a strong interest in monitoring and optimizing the quality of the services provided by their agents. Speaker classification can be employed for this purpose in a variety of ways.

For example, a classifier for the emotional state of the agents and/or the caller can be used to calculate numerical values that act as an indicator for the average quality of service provided over a certain period of time. Based on a large number of calls, even a classifier with a rather poor recognition rate on the utterance level is suitable to reliably detect relevant changes, such as an increasing number of angry callers or an increasing average stress level of the call center agents.

Speaker classification can also be employed to identify individual calls in a corpus of recorded call center conversations. For example, calls with angry users can be selected for the purpose of training agents to cope with this situation.

Telephone surveillance

Clearly, speaker classification methods offer the potential for increasing the effectiveness of telephone surveillance measures by government authorities, e.g. by preselecting calls conducted by certain subgroups of the population.

Influence on staff planning

There are studies showing that the success of a call center is higher (this could be measured in customer satisfaction or sometimes even in revenue) when the characteristics of the caller and the agent matches, i.e. they are in the same age range, equal social level, etc. So if it is possible to create a profile of the caller groups over time (in the morning young male high professionals call and in the afternoon elderly housewives from middle class families) it makes sense to try to match the structure of the call center agent groups in that time to the caller structure.

Cross- and Up selling

Imagine an automated ordering hotline where people call in and place their orders. What usually should be done in such an application is that the system gives the callers some proposals what should be ordered in addition (this sometimes is also done by human agents). If the system in that case could not only use the already ordered product as the base of the decision what to propose as up selling item but also the gender of the caller and the age and other speaker characteristics, the chance that the proposed additional item is ordered can be maximized. Just as an example: if someone orders a digital camera in an voice application, the up selling item for younger callers perhaps should be some software product like Photoshop for manipulating taken photographs whereas for elderly people an additional suitcase is proposed.

Quiz Games/Prize Competitions

In automatic prize competitions over the telephone the set of questions can be matched against the callers characteristic. Usually, the idea behind those quiz games is that callers have a quite good chance to get the right answers so that there is the chance for the company to establish a relation to the caller. If the systems is aware of teens calling in perhaps questions on currently successful pop bands is the right choice so that the callers have a quite good chance to know the right answers whereas for the older generation questions on classic music are more preferable.

Gaming, Fun

A related field of applications is given by applications in the gaming or entertainment sector. For example the love detector by Nemesysco Ltd. attempts to classify speech samples based on how much “love” they convey. Other entertainment application ideas comprise online-telephone role games or horoscope applications that variegate their prognoses based on some automatically detected speaker characteristic.

Conclusion

Generally spoken, the use of automated speaker classification is of high potential when it comes to the design of more natural interfaces in human machine communication. We introduced a taxonomy that might be helpful to think up possibilities for the integration of speaker classification in various application fields.

Furthermore we see from the large list of prototypical scenarios that can benefit from speaker classification strategies that this technology is indeed an important basis for enhancement in all kinds of application fields.

1 Acknowledgments

This work was partly funded by the EU in the framework of the NoE HUMAINE under grant IST-2002-507422, and by the German Federal Ministry of Education and Research (*BMBF*) in the framework of the SmartWeb project (Grant 01IMD01F). The responsibility for the contents of this study lies with the authors.

References

1. A. Batliner, F. Burkhardt, M. van Ballegooy, and E. Nöth, " A Taxonomy of Applications that Utilize Emotional Awareness ," in *Language Technologies, IS-LTC 2006* , T. Erjavec and J. Gros, Eds., Ljubljana, Slovenia, 2006, pp. 246–250, Infornacijska Druzba (Information Society).
2. R. Picard, "Affective Computing: Challenges," *Journal of Human-Computer Studies*, vol. 59, pp. 55–64, 2003.
3. F. Burkhardt, M. van Ballegooy, R. Englert, and R. Huber, "An emotion-aware voice portal," in *Proc. Electronic Speech Signal Processing ESSP*, 2005, pp. 123–131.