Estimating the Acoustic Context to Improve Speaker Classification

Christian Müller

German Research Center for Artificial Intelligence (DFKI),Saarbrücken
christian.mueller@dfki.de

Abstract. This paper describes, how an existing approach on speaker classification can be extended in a way that context is taken into account. First, the main issues of this approach are summarized. With respect to context classification, preliminary results of large corpus analyses are presented. It is outlined, how knowledge about the context can be applied to improve speaker classification.

1 Introduction

Speech as an input modality is getting more and more important especially for mobile applications. At the same time, designers are faced with the challenge of building systems, that are flexible enough for being useful within a wide range of situations and environments. Ideally, those systems should be able to adapt their behavior according to the special needs of the current user in the current situation (see [2]). To accomplish that goal, they need to have knowledge about what kind of user and what kind of situation they are dealing with. Within the project M3I, we are investigating the possibility to derive this knowledge directly from the characteristics of the speech.

In a former phase of the project, we developed a speech-based approach on user classification (see [5]). In particular, the pilot system estimates the speakers' gender and age class ($<= 65$, > 65). The approach is two-layered: on the first layer, low level features are extracted from the speech sample and processed by multiple pattern classifiers [1] (e.g Artificial Neural Networks, [4]). On the second layer, we combine their results and treat the inherent uncertainty using a Dynamic Bayesian Network. Since context classification now plays an important role in that phase, the second layer is described in more detail in section 3.

Based on findings from the literature about vocal aging (see e.g. [3]), we discriminate acoustic features (e.g. pitch, jitter, shimmer) and prosodic features (e.g. speech pauses, articulation rate). This poses the following trade-off: On the one hand, the prosodic features are less predictive and harder to extract than the acoustic ones. On the other hand, the acoustic features are more sensitive to noise, i.e. the classification accuracy decreases when there's noise in the background. Therefore it is desirable to know whether there is noise or not.

2 Context classification

Speech – especially in conjunction with mobile systems – is usually not laboratory-like but contains portions, that do no belong to the actual utterance. We refer to these portions as "context" (instead of "background noise"), because they provide information about the situation in which the user is interacting with the system. For the classification of the context, we consider the following features to be relevant: the Minimal-to-Average-Intensity-Ratio (MIN2AVG) and the Noise-to-Harmonicity-Ratio (N2H). The former measure indicates, how loud the most quiet sequence is compared to the average. It is based on the assumption, that context-free speech always contains (at least tiny) pauses which constitute portions of sound with a very low intensity. The latter, N2H, measures the proportion of the harmonic elements with respect to non-harmonic ones. Speech mostly consists of harmonic elements (particularly the vowels) whereas most kind of context is non-harmonic (noise).

We conducted a series of analyses on a large "artificial" corpus. To build this corpus, we separately recorded several kind of contexts and overlayed them with background-free speech samples. In detail, we recorded samples from a crossroad, a highway, a jack-hammer, voices, a book-store and a public library. Each sample was overlayed with 6300 utterances, that contained no background noise using seven different "overlay-factors". These factors correspond to the loudness (intensity) of the context. Thus, our artificial corpus contains 6300 utterances * 6 contexts * 7 overlay factors = 264.600 samples.

Figure 1 summarizes the results of N2H and MIN2AVG analyses. As it is shown, we can identify three categories of contexts: "quiet" contexts with high N2H values and low MIN2AVG values; "noisy" contexts (e.g. crossroad, highway and jack-hammer) with low N2H values and high MIN2AVG values and finally "voicy" contexts (e.g. voices, bookstore, and library) with both high N2H and MIN2AVG values.

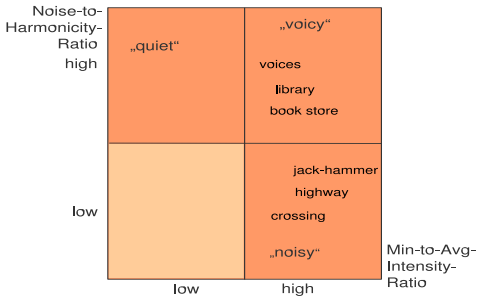

Fig. 1. *Grouping of contexts based on Noise-to-Harmonicity-Ratio and Minimal-to-Average-Intensity-Ratio*

By the time when this contribution was written, we finished training and testing of classifiers for two context categories – quiet and noisy. We conducted a ten-fold cross-validation[1] which yielded in a true positive rate (TPR) of 0.99 for both classes, i.e. "quiet" samples where classified as quiet and "noisy" samples as noisy in 99% of all cases. Although the simplifications we made were rather restrictive, these are promising results, that can already be reasonably applied as it is shown in the following section.

3 Improving Speaker Classification

As already pointed out in section 1, in M3I causal relationships are modeled on the second layer using Bayesian Networks (BNs, see [6]). A BN consist of two parts: (a) a directed acyclic graph G, that encodes the causal relationships between the considered random variables and (b) a set of conditional probability tables (CPTs), that quantify the uncertain relationships represented by the links of G. Figure 2 shows the structure of the BN, that we used to solve our classification task. Let us first look at the speaker classification part (left hand side): There are direct causal dependencies between the speaker's actual gender/age and the results produced by the classifiers for a particular speech sample. Analogously, there are direct causal dependencies between the actual level of noise and the results produced by the corresponding classifier.

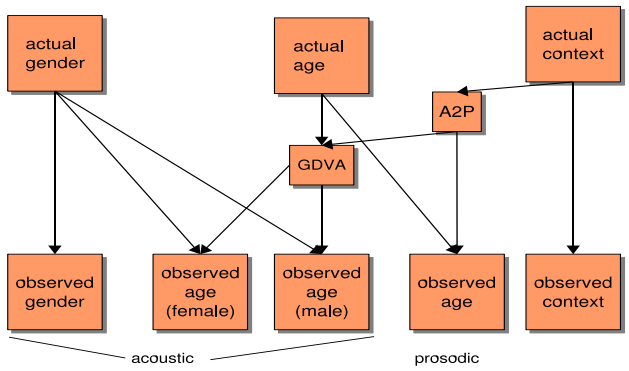

Fig. 2. *Modeling causal relationships on the second layer using Bayesian Networks.*

The second part of the BN, the conditional probabilities, can be computed as the true positive/negative rates of the learned classifiers when applied to

[1] With this evaluation method, the data is split into ten approximately equal partitions. Each one is used in turn for testing while the remainder is used for training. The procedure is repeated ten times.

corresponding samples. For example, the probability, that the context classifier indeed classifies a noisy context as noisy, can be estimated on the basis of the classifier's accuracy (in this case 0.99).

Figure 2 includes an example on how the DBN is used to acquire knowledge about the speaker as well as the context. The results of the classifiers are used as evidences for the variables at the figure's bottom, that are interpreted by applying the BN reasoning mechanism. This procedure yields posterior probabilities (conditioned on the basic classification results) for the states of all three variables of interest on the figure's top line — gender (female/male), age (elderly/non-elderly) and context (quiet/noisy).

Besides being a "variable of interest", context influences speaker classification, since there's a connection via the node A2P-balance: when the context is classified as quiet, the acoustic classifiers ('observed gender', 'observed age (male)', and 'observed age female') have a larger influence to the predicted age than the prosodic ('observed age'). With a noisy context it is vice versa. In this vein, the information about the context helps solving the above formulated trade-off.

4 Summary

In this paper, we outlined, how an existing approach on speaker classification can be extended in a way that also context is taken into account. We presented preliminary results of large corpus analyses that were used – in a simplified form – to train a classifier which estimates on a very high level of accuracy, whether a given speech sample was recorded in a quiet or in a noisy environment. We also pointed out, how we integrated this classifier into our approach on on speaker classification, and how in this way the preference trade-off between acoustic and prosodic features can be resolved. A system demonstration will be available at the conference.

References

1. Richard O. Duda and Peter E. Hart. *Pattern Classification (2nd Edition)*. Wiley, New York;London;Sydney, 2001.
2. Jorge Joaquim. Adaptive tools for the elderly: new devices to cope with age-induced cognitive disabilities. In *Proceedings of the 2001 EC/NSF workshop on Universal accessibility of ubiquitous computing*, pages 66–70. ACM Press, 2001.
3. Sue Ellen Linville. *Vocal Aging*. Singular, San Diego, Ca, 2001.
4. Tom M. Mitchell. *Machine Learning*. McGraw-Hill Series in Computer Science. McGraw Hill, London;New York;Paris, 1997.
5. C. Müller, F. Wittig, and J. Baus. Exploiting Speech for Recognizing Elderly Users to Respond to their Special Needs. In *Proceedings of the Eighth European Conference on Speech Communication and Technology (Eurospeech 2003)*, 2003.
6. Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.