Towards a Multilingual Approach on Speaker Classification

Christian Müller, Michael Feld

German Research Center for Artificial Intelligence

{Christian.Mueller, Michael.Feld}@dfki.de

Abstract

This paper outlines a framework for a multilingual speaker classification system which is based on an underlying language identification module. First, the AGENDER speaker classification technology is introduced, a two-layered approach which primarily recognizes the speakers' age and gender but also incorporates novel domain-independent aspects that can be applied to other speaker characteristics like emotions or cognitive load. Then, it is pointed out that one of its major drawbacks consists of the fact that it has not been verified that the chosen set of speech features also works for other languages, especially for those with different phonological aspects. To overcome this drawback, it is suggested to extend AGENDER with a language identification module. The module presented here is designed to meet the requirements of a specific telephone-based application (which itself is not within the focus of this paper): The languages German, English and Turkish shall be discriminated on the basis of the initial utterance of the speaker; for each of the possible languages, hypotheses about the nature of the initial utterance are available; the domain encompasses a list of English product names. Although the suggested method is as yet only partly implemented, the first evaluation results are very promising: Turkish could be identified with an accuracy of 71.75 %, German with an accuracy of 78.39 %, and English with an accuracy of 79.89 %. Besides this, the paper outlines the use of the language identification module within a multilingual version of AGENDER.

1. Introduction

In our previous work, we described the AGENDER speaker classification technology, a two-layered approach which primarily recognizes the speakers' age and gender, but also incorporates novel domain-independent aspects that can be applied to other speaker characteristics like emotions or cognitive load. Due to its classification accuracy [1, chap. 8], its flexible way of fusing the results of multiple classifiers [1, chap. 9] as well as its multiple-platform architecture [2], the project is regarded as very successful, attending e.g. vital interest from telecommunications industry.

Today, one of AGENDER's major drawbacks consists of the fact that it has not been exhaustively investigated, whether the approach is language-independent or not. This paper outlines our attempt to overcome this drawback. Particularly, we present a framework for a multilingual speaker classification system which is based on an underlying language identification module. This framework is designed to meet the requirements of a specific telephone-based application which itself is developed in our lab and therefore within the focus of this paper: On the basis of the initial utterance, the caller is classified according to her/his age and gender as well as her/his nationality. It is assumed that the caller is speaking in her/his mother tongue. The number of possible languages is restricted to three, namely German, English, and Turkish. The application scenario is closed-world, i.e. domain-specific hypotheses about the most likely candidates for the initial utterances are given for each language. Finally, the domain encompasses a list of English product names – a fact that should be taken into account during the language identification process.

The remainder of this paper is organized as follows: Section 2 summarizes the major aspects of the existing AGENDER system. Section 3 describes our approach on the language identification task. Section 4 outlines the issues of MULTILINGUAL AGENDER.

2. The Agender Approach

The AGENDER approach on speaker classification represents a combination of data-driven and knowledge-based aspects. The models are built on the basis of data stemming from extensive empirical analyses. In the current version, the age classes are defined as follows: The class CHILDREN represents speakers up to and including an age of 12 years. The class TEENAGER encompasses speakers between 13 and 19 years. Speakers between 20 and 64 years belong to the class (younger) ADULTS. The class of SENIORS begins with 65 years. Hence, in conjunction with the gender, the classification task consists of a total of eight classes. The speaker classes are denominated with one of the capital letters C,Y,A or S representing the age class, followed by one of the lower case letters m or f representing the gender.

The overall corpus used in the underlying empirical studies consists of three parts: the German corpus BAS [3], the English corpus Timit [4] and an English corpus that was provided by Nuance[1] for this purpose. The set of features was assembled on the basis of literature studies: (1) pitch, the speaking fundamental frequency; (2) jitter and shimmer, i.e. microvariations of the F0-frequency and amplitude. Both features were measured with multiple algorithms including RAP and PPQ for jitter and APQ3 and APQ11 for shimmer [5]; (3) the harmonics-to-noise-ratio which quantifies the relative amount of additive noise in the voice signal;(4) the articulation rate;(5) the number of speech pauses; (6) the duration of speech pauses.

In AGENDER, those phases of pattern recognition which correspond to the feature extraction and classification, are called the *first layer*. With respect to the classification, the following well known machine learning methods have been investigated: 1. Naive Bayes (NB), 2. Gaussian Mixture Models (GMM), 3. k-Nearest-Neighbor (KNN), 4. C 4.5 Decision Trees (C45), 5. Support Vector Machines (SVM) and 6. Artificial Neural Networks (ANN)[2].

[1] http://www.nuance.com (2006/02/10).

[2] Multilayer Perceptron Networks

8-class-problem						total accuracy 63.50 %		
	Cf	Cm	Yf	Ym	Af	Am	Sf	Sm
Cf	**76.09**	4.07	13.6	5.06	0.54	0.05	0.44	0.15
Cm	54.25	**12.37**	12.52	15.51	1.13	0.25	3.78	0.2
Yf	54.15	2.41	**27.44**	13.16	1.28	0.1	1.37	0.1
Ym	20.08	3.98	6.33	**59.25**	1.03	1.13	4.96	3.24
Af	0.25	0	0.2	0.54	**84.73**	3.44	6.92	3.93
Am	0	0	0	0.74	3.53	**87.87**	1.57	6.28
Sf	0.59	1.13	0.15	2.5	3.78	0.93	**77.07**	13.84
Sm	0	0.05	0	1.67	1.18	1.47	12.47	**83.16**

Table 1: Confusion matrix for the 8-class-problem with an ANN. The total accuracy is 65.50 % with a chance level of 12.5 %.

	Af	Am
Af	**90.63**	9.37
Am	4.36	**95.64**

Table 2: Confusion matrix for the gender recognition problem with an ANN. The total accuracy is 93.14 %.

	CfCmYfYmAfAm	SfSm
CfCmYfYmAfAm	**92.24**	7.76
SfSm	3.02	**96.98**

Table 3: Confusion matrix for the discrimination of seniors from all other age classes with an ANN. The total accuracy is 94.61 %.

The results are very promising: The classification accuracy of all methods in the test were significantly higher than the chance level. Table 1 shows a confusion matrix of the best-performing method ANN. The columns represent the actual speaker class and the rows the results of the classifier. Hence, the diagonal (bold numbers) contains the correctly classified cases, the so called true positive rates (TPRs). The values are percentages that were calculated via ten-fold cross validations.

The overall accuracy for the eight-class problem that was obtained with the method ANN is 64.5 % which is five times better than the chance level (12.5 %). With TPRs between 77.07 and 87.87 %, the accuracies for adults and seniors are very satisfying, while – on the first look – those for the remaining speaker classes (except Cf) are not. The confusion matrix however shall not only be interpreted in terms of TPRs; it is likewise important to consider the distribution of the misclassified cases. The majority of misclassified Cm for example has been categorized as Cf, a fact that absolutely conforms with our hypotheses. In general, most of the confusion occurred within consecutive cells whereas "long distance" confusions (indicating noisy classifiers) occurred rather seldom. This interpretation is supported by the high accuracy of 94.61 % (1.89 times chance level) provided in table 3 where the age classes are grouped in a way that seniors are discriminated from all other classes. With respect to a pure gender estimation, a likewise high accuracy of 93.14 % (1.86 times chance level) was achieved (see table 2).

Figure 1 compares the performances of the various classification methods. The x-axis represents the total accuracy (average TPRs) and the y-axis the balance (based on the standard deviation of the TPRs). With an overall accuracy of 64.5 % for the eight-class problem (with a chance level of 12.5 %) the neural network (ANN) performed best, followed the k-nearest-neighbor model (KNN). The rather simple decision tree method (C 4.5) also performed surprisingly well, especially with respect to the balance of the TPRs. The parametric methods Gaussian Mixture Models (GMM) and Naive Bayes (NB) however fell short of the expectations. This is possible due to the fact that the GMM implementation used in this test did not learn the weight vector using the EM-algorithm but estimated it on the basis of an upstream evaluation [1].

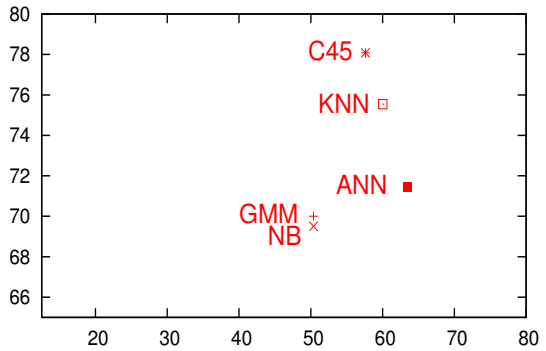

Figure 1: Comparison of various classification methods. X-axis: total accuracy (average true positive rates). Y-axis: balance (based on the standard deviation of the true positive rates).

Besides these positive results, the AGENDER speaker classification approach distinguishes itself by means of a special post processing technique, the so called *second layer*: Multiple post processing problems are solved with one single mechanism, namely dynamic Bayesian networks (DBNs). [1] provides examples on how DBNs can be used for: 1. explicitely modeling the classification inherent uncertainty; 2. incorporating top down knowledge into the decision making process, like e.g. the fact that depending on the context, certain classifiers are more reliable than others; 3. fusing the results of multiple classifiers with respect to one utterance (static fusion) as well

as several consecutive utterances (dynamic fusion).

One major drawback of the current version of AGENDER consists of the fact that mixed-language material containing both German and English speakers was used for training and evaluation of the classifiers. It has yet to be verified that the chosen set of speech features also works for other languages, especially for those with different phonological aspects. Furthermore, it needs to be tested if classifiers trained with speakers of a single language perform better on utterances in that specific language. To incorporate language-specific classification into the AGENDER approach, the automatic identification of a speaker's language is needed.

3. Automatic Language Identification

As described above, the requirements of our language identification module are the following. First of all, three languages shall be discriminated, namely German, English, and Turkish. Secondly, the classification should be done on the basis of the initial utterance of the speaker. For each of the possible languages, hypotheses about the nature of the initial utterance are available. Finally, it should be taken into account that the domain encompasses a list of English product names.

In contrast to Turkish, the discriminability of English and German is well documented in the literature. In a test with a total number of seven languages, [6] obtained a classification accuracy of 77.1 % (English) respectively 75. % (German) using a phonotactic model with phoneme strings that were automatically derived by an so called *Ergodic HMM* (EHMM). With a parallel syllable-like unit recognition method, [7] obtained in a test with eleven languages an accuracy of 80 % - 85 % for English and 65 % - 85 % for German, depending on the length of the test utterances (10s respectively 45s).

Although not yet exhaustively investigated, it can be assumed that Turkish can likewise be discriminated from the two Germanic languages. In contrast to German and English, Turkish is a phoneme-based language like Finnish or Japanese [8]. There exists nearly a one-to-one mapping between written text and its pronunciation. It is much different from Indo-European languages in that its morphology is agglutinative and suffixing [9]. The Turkish vowel inventory is small and very symmetric; its eight phonemic vowels are grouped into foursomes with respect to the features of height, backness and rounding. There are no diphthongs in the language, and all vowels of the native vocabulary in Turkish language are phonemically short. There is also only a small number of consonants, and consonant clusters within words are not allowed at all. In case of proper names and borrowed words, vowels are usually lengthened and consonant clusters broken up by native speakers, inserting additional vowels according to the vowel harmony rules.

Figure 2 describes the general approach on language identification that we pursue in the MULTILINGUAL AGENDER project. The box in the center of the diagram describes the so called phonotactics model (PM). It is motivated by a method that [10] calls PRLM (phone recognition followed by language modeling). In this language-id approach, for each language l from the set L of languages to be identified 1) training messages are tokenized by a single-language phone recognizer; 2) the resulting symbol sequence associated with each of the training messages is analyzed; 3) n-gram probability distribution language model is estimated. The important aspect is that the n-gram model is trained from the output of the phone recognizer, not from orthographically or phonetically labeled data. During the recognition phase, a test message is tokenized and the likelihood that the resulting symbol sequence was produced by language l is calculated for each $l \epsilon L$. The language that belongs to the model with the highest likelihood is selected as the language of the message. Note that this approach can employ a single-language phone recognizer trained from speech in any language. Although it is desirable to possess a phone recognizer which doesn't incorporate any language-specific constraints *during* the Viterbi decoding (unconstrained recognizer), for bootstrapping any reliable single language phone recognizer should be suitable. In MULTILINGUAL AGENDER we use the *Sphinx 2* recognizer from CMU (see http://www.speech.cs.cmu.edu/), because it is well suited for integration into the existing platform. A PHP script that runs *Sphinx* in *allphone* batch mode on a speech corpus is used to generate a training file for each language. The corpora we used to train these models was the *Timit* corpus for English [4], and the *GlobalPhone* corpus for Turkish and German [11]. The training file contains a list of phones for each utterance in the corpus including time information. For now, only the phone string is used to train the language model; pauses and segment length information are ignored. The actual language model is composed of a table of weighted probabilities $\tilde{P}$ for every bigram that occurs in the language. Again a PHP script is executed to compute the $\tilde{P}$ table for all languages. The computation is performed according to the formula suggested in [10], which takes into account the weighted distribution of both bigrams and unigrams within a language through n-gram histograms. For bigrams that did not occur in any training sample, a $\tilde{P}$ of 0 is assumed. To classify a new utterance, our embedded implementation first converts the audio to a phone sequence using the *Sphinx 2* library. Afterwards, the classifier determines the likelihood $\Lambda(l)$ for each language l according to

$$\Lambda(l) = \sum_{t=1}^{T-1} \log \tilde{P}_l(w_t)$$

where T is the length of the sequence, w_t is the biphone at position t and $\tilde{P}_l(b)$ is the value of $\tilde{P}$ for biphone b in the language l.

The box on top of the diagram represents the so called pseudo-syllable model (PSM) which is based on the cv-structure of the speech and therefore requires an upstream cv-segmenter. On the basis of the cv-structure, a set of features is calculated that is reported to be significant for the language identification task [12]: the average proportion of vocalic intervals (%V), the average standard deviations of consonantal intervals (ΔC) and the average standard deviations of vocalic intervals (ΔV). For cv-segmentation the output of the phone recognizer is categorized as consonantal or vocalic. In addition, it is annotated with the length of the segment because this information is needed for the calculation of the above described features. The box at the bottom of the diagram finally described the so called word spotting model (WSM). The WSM directly takes into account the characteristics of the telephone-based application: 1. Since hypotheses are available about the nature of the initial utterance, it is reasonable to spot a list of expected words for each language. We call these lists *positive lists* (+ lists). If a word from a positive list of a certain language l is spotted, the likelihood of l is increased. The *negative list* (- list), on the other hand side, corresponds to the set of English product names. If one or more words from that list is spotted, the certainty value of the language-id module is extenuated.

As indicated in figure 2, the MULTILINGUAL AGENDER ap-

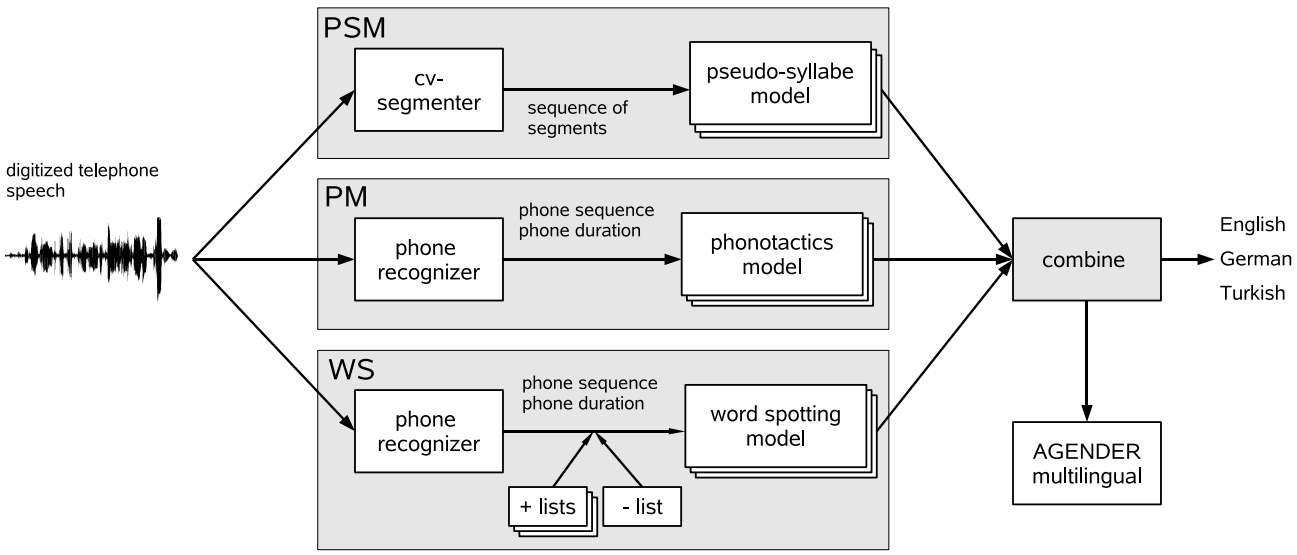

Figure 2: Overview of selected methods for language identification and their integration into AGENDER

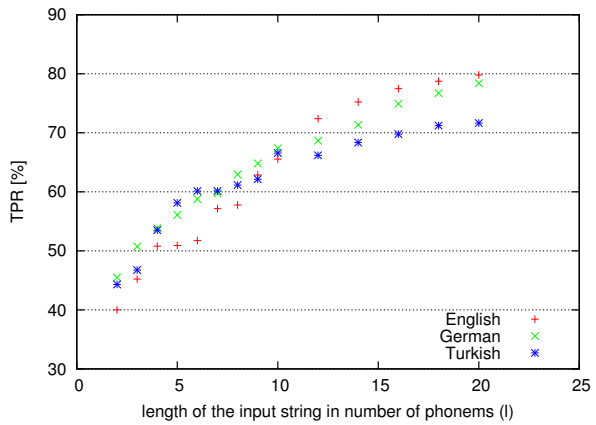

Figure 3: First evaluation results of the language identification module.

proach envisages a combination of the different classifiers to finally identify the language. However, at the moment only the phonotactics model is implemented and integrated into the AGENDER system. In figure 3, the first evaluation results of the module are presented. The x-axis represents the length of the test utterance in number of phonems (l) and the y-axis the true positive rates in percent (TPR). Each of the languages Turkish, German, and English is drawn separately. As expected, the TPRs rise as a function of the l and tend to a maximum value when l tends to infinity. The maximum value for l in the test was restricted to 20 due to the limited length of the utterances in the corpus. Here, an accuracy of 71.75 % for Turkish (695 test samples), 79.8 % for English (630 test samples), and 78.39 % for German (1009 test samples) was obtained. Note that $l = 20$ corresponds to an utterance length of only 1.88 seconds on an average. We expect even better classification accuracies when all three methods depicted in figure 2 are implemented.

4. Agender in a Multilingual Setting

Following the approach outlined in 3, we can extend the set of classifiable properties to the language of an utterance which is by itself a requirement for the telecommunication application. However, language identification is far more interleaved with the other parts of the AGENDER system and can help us to answer some questions and offers possibilities to improve existing methods.

The standard version of AGENDER was trained with corpuses containing speakers of different languages. The high accuracy of the system suggests that the AGENDER concept can be applied to all of the target languages, but nonetheless we need to obtain a deeper understanding of the gender-specific vocal aging process in each of the languages to reinforce this hypothesis. We also expect further empirical studies in this field to indicate that cultural differences can have an effect on the speech features used for classification of age and gender. Even if that effect will be minor, it would imply that a language-specific classifier may perform better than a generic (language-independent) one, given that the language was detected correctly. Hence, our task will include building specialized classifiers that are trained only with English, German and Turkish speakers, respectively. We will then conduct three types of evaluation: (1) using language-specific classifiers on utterances in the same language for which the classifier was trained, (2) using a classifier for one specific language on utterances in each other language, and (3) using the generic classifier from standard AGENDER to classify speakers form each language separately. By incorporating the results of these evaluations, we hope to learn more about *when* language-specific classifiers for age and gender perform differently than the language-independent version and *why*.

Following these studies, the language identification methods will be used to extend and improve AGENDER, putting it into a multilingual context. By taking advantage of the flexibility and extensibility of the current architecture, we can integrate both language identification and language-specific speaker classification into the existing system with little effort to create MULTILINGUAL AGENDER.

5. References

[1] C. Müller, "Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht [Two-layered Context-Sensitive Speaker Classification on the Example of Age and Gender]," Ph.D. dissertation, Computer Science Institute, University of the Saarland, Germany, 2005.

[2] M. Feld, "Erzeugung von Sprecherklassifikationsmodulen für multiple Plattformen," Master's thesis, Fachbereich 6.2 Informatik, Universität des Saarlandes, Deutschland, 2006.

[3] F. Schiel, "Speech and Speech-Related Resources at BAS," in *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, 1998, pp. 343–349.

[4] J. e. A. Garofolo, *DARPA TIMIT CD-ROM: An Acoustic Phonetic Continous Speech Database*. Gaithersburg, MD, USA: National Institute of Standards and Technology, 1998.

[5] R. Baken and R. Orlikoff, *Clinical Measurement of Speech and Voice*, 2nd ed. San Diego, Ca, USA: Singular Publishing Group, 2000.

[6] P. Matejka, I. Szke, P. Schwarz, and J. Cernocky, "Automatic Language Identification using Phoneme and Automatically Derived Unit Strings," *Lecture Notes in Computer Science*, vol. 2004, no. 3206, p. 8, 2004.

[7] N. T. and H. Murthy, "Language identification using parallel syllable-like unit recognition," in *Proceesings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'04)*, vol. 1, Montreal, Canada, 2004, pp. 401–404.

[8] Ö. Salör, B. Pellom, T. Ciloglu, K. Hacioglu, and M. Demirekler, "On developing new text and audio corpora and speech recognition tools for the turkish language," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 02)*, Denver, Colorade, USA, 2002, pp. 349–352.

[9] K. C̆arki, P. Geutner, and T. Schultz, "Turkish LVCSR: Towards Better Speech Recognition for Agglutinative languages," in *Proceesings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'00)*, Istanbul, Turkey, 2000, pp. 3688–3691.

[10] M. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, 1996.

[11] T. Schultz and A. Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, vol. 1, Rhodos, Greece, 1997, pp. 371–373.

[12] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, 1999.