Multilingual Speaker Age Recognition: Regression Analyses on the Lwazi Corpus

Michael Feld [*1], Etienne Barnard [#2], Charl van Heerden [#3], Christian Müller [*4]

[*] *German Research Center for Artificial Intelligence, Germany*
[1] michael.feld@dfki.de [4] christian.mueller@dfki.de

[#] *Human Language Technology Research Group, Meraka Institute, CSIR, South Africa*
[2] ebarnard@csir.co.za [3] cvheerden@csir.co.za

***Abstract*—Multilinguality represents an area of significant opportunities for automatic speech-processing systems: whereas multilingual societies are commonplace, the majority of speech-processing systems are developed with a single language in mind. As a step towards improved understanding of multilingual speech processing, the current contribution investigates how an important para-linguistic aspect of speech, namely speaker age, depends on the language spoken. In particular, we study how certain speech features affect the performance of an age recognition system for different South African languages in the Lwazi corpus. By optimizing our feature set and performing language-specific tuning, we are working towards true multi-lingual classifiers. As they are closely related, ASR and dialog systems are likely to benefit from an improved classification of the speaker.**

In a comprehensive corpus analysis on long-term features, we have identified features that exhibit characteristic behaviors for particular languages. In a follow-up regression experiment, we confirm the suitability of our feature selection for age recognition and present cross-language error rates. The mean absolute error ranges between 7.7 and 12.8 years for same-language predictors and rises to 14.5 years for cross-language predictors.

I. Introduction

Speech has both universal aspects, on the one hand, and language- or culture-specific aspects, on the other. Despite centuries of theoretical debate [1], the balance between these two classes of properties is still highly controversial. As a practical matter, it is important to improve our understanding of this balance both in order to develop multi-lingual speech-processing systems and to utilize cross-language sharing so that the number of languages for which speech technologies are available can be expanded.

At a superficial level, it is clear that spoken language differs across cultures and languages along a multiplicity of dimensions, ranging from acoustic phonetics through grammar, vocabulary and metaphor to pragmatics and discourse strategies. (Some of these differences, such as those involving metaphor or acoustic phonetics, may be pronounced even for cultural groups that share the same language, whereas other factors such as grammar tend to be more widely shared by speakers of the same language.) However, the effects of culture on the various facets of speaker classification have to date received comparatively little attention. Various authors have reported that modern approaches to speaker identification are reasonably insensitive to the language being spoken (see, e.g., [2]); although statistically significant differences in the performance of speaker-verification algorithms on different languages from the same corpus have been reported [3], these differences are relatively small in magnitude. Emotion recognition, on the other hand, has been shown to depend strongly on the language being spoken [4].

For the case of age classification, which we consider in the current paper, we are not aware of any cross-cultural or multi-lingual studies, although there are a lot of useful ASR related applications associated with this task. Age (combined with gender) can for example be used to adapt the ASR system to an individual speaker. Furthermore, for interactive voice response systems, waiting music can be adapted, age dependent advertisements can be presented to callers in the waiting queue, or speaking habits of the text-to-speech module can be changed. Statistical information on the age distribution of a caller group is also of interest for the provider [5].

In order to extend this technology from a mono- to a multi-lingual setting, we have set out to investigate the influence of language on an initial feature set that has been employed for age classification. In a next step, we hope to be able to find out in how far our current speaker classification approach – the combination of features, classification architecture and pre-trained models – is dependent on language. In Section II we summarize the multilingual corpus that was used for our experiments – the Lwazi automatic speech-recognition (ASR) corpus – as well as some pertinent facts about the languages contained in that corpus. Section III contains a description of the feature sets used during the analysis, together with a statistical analysis of some of the features across languages. A comparative evaluation in terms of classification errors is provided in Section IV, and Section VI summarizes our overall conclusions.

II. The Lwazi ASR corpus

The Lwazi ASR corpus was developed as part of a project that aims to demonstrate the use of speech technology in information service delivery in South Africa [6], [7]. In particular, the three-year Lwazi project (2006-2009) produced the core tools and technologies required for the development of multilingual voice-response systems in all eleven of South

Africa's official languages, and piloted the use of these technologies in government information service delivery.

The Lwazi ASR corpus consists of annotated speech data in the languages listed in Table I, which also summarises the amount of speech available in each language. This data was collected in South Africa over the telephone, by soliciting callers in each of the languages from a variety of backgrounds. Approximately 100 male and 100 female first-language speakers contributed speech in each of the languages, and an approximate balance between mobile and fixed-line telephones was maintained across languages. This corpus was restricted to adult speech; details of the distribution of speaker ages are provided in Section III. For some languages, extensive dialectal variation exists within South Africa. However, this variation is not well documented; for the purposes of the corpus, the intent was to concentrate on the dominant dialects of each languages, but dialects were not rigorously controlled.

TABLE I
THE OFFICIAL LANGUAGES OF SOUTH AFRICA, THEIR ISO 639-3:2007 LANGUAGE CODES, AND THE AMOUNT OF SPEECH CONTAINED IN THE LWAZI CORPUS

Language	code	# total minutes	# speech minutes
isiZulu	zul	525	407
isiXhosa	xho	470	370
Afrikaans	afr	213	182
Sepedi	nso	394	301
Setswana	tsn	379	295
Sesotho	sot	387	313
SA English	eng	304	255
Xitsonga	tso	378	316
siSwati	ssw	603	479
Tshivenda	ven	354	286
isiNdebele	nbl	564	465

The languages in Table I fall into two broad families, with Afrikaans and English being Germanic languages and the remaining nine languages belonging to the Bantu family of languages (in particular, the Southern Bantu sub-family). The co-location of these widely different groups of families in the same country is a historical accident; although their many years of co-existence have lead to some mutual influences, the two groups remain separated by a wide linguistic gulf. For example, the Bantu languages of South Africa are tonal languages characterized by an extensive system of noun classes; they are strongly agglutinative, with affixes playing a variety of syntactic and semantic roles; their syllables tend to have regular CV or V structures. In all these respects the Southern Bantu languages differ from English and Afrikaans, which are fairly typical Germanic languages. Hence, these languages are a good testing ground to search for differences in the way that speaker age is expressed in speech.

III. CORPUS ANALYSIS

With over 14 GB of speech data, the Lwazi corpus is quite substantial and working with it requires significant computing time and processing power. Also, it was developed under developing-world conditions, where limitations in infrastructure and the availability of skilled personnel are expected to impact on corpus quality. An initial random listening and signal analysis was therefore undertaken by one of us (MF); it revealed some interesting facts about the material. Compared to widely-used speech corpora such as GlobalPhone[8] or Timit[9], there is considerable background noise, which is a consequence of the fact that many speakers were speaking on mobile telephones and from everyday locations. For the same reasons, the amplitudes of the speakers are much more variable than in standard corpora. On the signal level, the data contained varying DC offset and even some clipping on some speakers. Again, that is explained by the absence of constant recording conditions with respect to the sender's microphone, the line and the recording equipment on the receiver's side. All of this is part of the compromise when trying to find a large number of native speakers for these languages, and it has to be taken into account when comparing the results with evaluations done on other corpora. Note that the effects were mostly randomly distributed over all languages - there were no visible artifacts restricted to a single language. Thus, for a comparative study on classification performance, these observations are not considered critical.

To get an initial idea of how the influence of language and culture manifests itself in a speaker's voice, a semi-automated corpus analysis was performed. In general, rather than simply taking a random collection of features and processing them with various out-of-the-box classification algorithms, it is more purposeful to take a look at the expressiveness of some of the available features individually (as was also done in [10]). This not only saves time, but also gives a better understanding of the decision criteria and simplifies the task of fine-tuning the classifier later on. A representation that is well suited for this purpose is the Gaussian approximation of the distribution of feature values. Sketched over all utterances in the corpus or a particular language, it provides a graphical comparison of the differences between the target classes, i.e. ages and genders in our case. This task can be automated to some extent, but many of the more interesting relations are hard to recognize by a machine and can usually only be spotted by manually looking at the results. In order to make the results more comparable, we chose the same definition of classes that had been used to train our previous classifiers (see Table II). For now, we restricted our experiments to a subset of languages as the age labels were not yet fully available for the remaining languages. An age histogram can be found in Figure 1. As can be seen, there are almost no children in the corpus and only a relatively low number of elderly people. Consequently, the results given in Section V do not consider the *children* class and the error rates may not be expressive enough to yield reliable benchmarks for seniors in general.

The features that were selected for the corpus analysis are acoustic long-term features computed on full utterances that have proven useful already in the past [10]. They are derivatives of the pitch, jitter and shimmer families of features computed as averages on whole utterances and were obtained using *Praat*[11]. This decision was made in spite of our recent findings, where MFCC features generally produced lower error

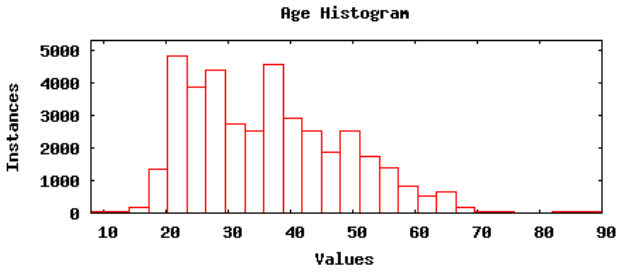

Fig. 1. Distribution of ages for a subset of seven languages in the Lwazi corpus.

TABLE II
CLASSES USED FOR THE CURRENT AGE/GENDER SPEAKER CLASSIFICATION SYSTEM

Index	Class	Ages
1	Children	7 – 14
2	Young female	15 – 24
3	Young male	15 – 24
4	Adult female	25 – 54
5	Adult male	25 – 54
6	Senior female	55 – 80
7	Senior male	55 – 80

rates than explicit long-term features [12]. However, long-term features are still the preferred way to get an initial performance measurement because they are much more expressive to humans than raw MFCC coefficients, and thus provide a better understanding of the phonetic causes for speakers of some languages being classified better than those of others. Also, there are without doubt certain overlaps concerning the information the two feature sets contain.

From the data we were looking at, a large number of distributions can be examined, comparing either languages grouped by age class or age groups across language. In many of these charts, there is indeed a notable difference in the distribution of values for the individual languages. A selection of these graphs is provided in this section in order to illustrate some of the most interesting cases where major deviations are visible. Two of the features with a very characteristic language-specific average are the mean and standard deviation of pitch. Figure 2 shows the typical distribution for speakers of South African English, with male speakers having generally lower pitch than female speakers. The *adult female* voices are a bit higher than expected, which is probably due to the slightly uneven distribution of ages (bias towards < 30). It confirms that pitch is a good feature for gender recognition, and to some extent helpful to distinguish ages. In order to see how language affects this circumstance, we next study that feature for the individual languages and a specific age class. In Figure 3, this was done for *young female* speakers of isiZulu and Sepedi. Our analyses revealed that voices of isiZulu speakers are on average 25 Hz lower than those of Sepedi speakers, which would make them easily confusable with seniors if the system was trained only on data from

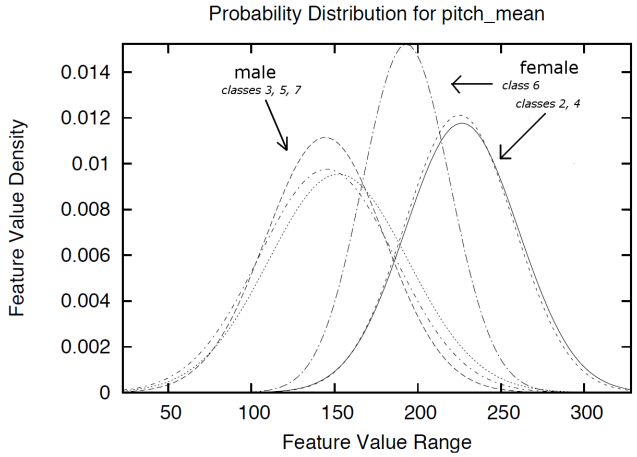

Fig. 2. Feature value distribution of mean pitch over the age/gender classes for South African English speakers. The curve peaks represent the class mean while the width indicates its inner-class variance. This Figure shows the typical separation of female and male voices. There were no voices of children in the data.

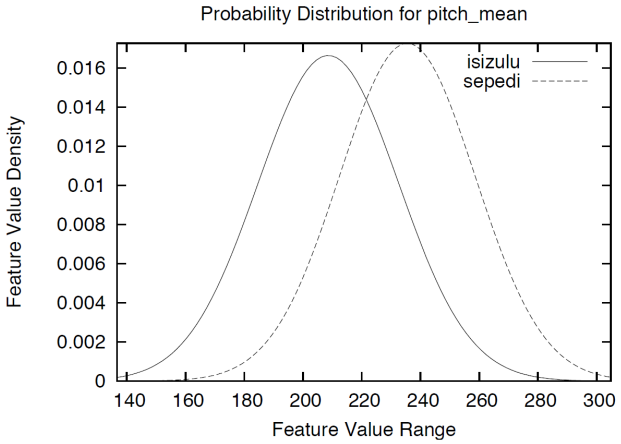

Fig. 3. Comparison of the mean pitch distributions for isiZulu and Sepedi of speakers in the *young female* class.

Figure 2. Figure 4 shows a similar behavior for the standard deviation of pitch with different speakers.

The observation that some languages are more different than others in terms of long-time features surfaces in several of the charts, and the actual ordering of languages changes depending on the feature that is considered. The Germanic languages in the Lwazi corpus are usually rather close. Although these observations are at first an obstacle for speaker classification, if such an aspect is stable over a set of languages, it can be exploited by creating language-specific classifiers. For example, a particularly low jitter (micro-variations in the pitch level) can be observed for adults speaking Sepedi (see Figure 5), while a high shimmer (micro-variations of the amplitude) appears to be characteristic for adult female Zulu speakers (see Figure 6).

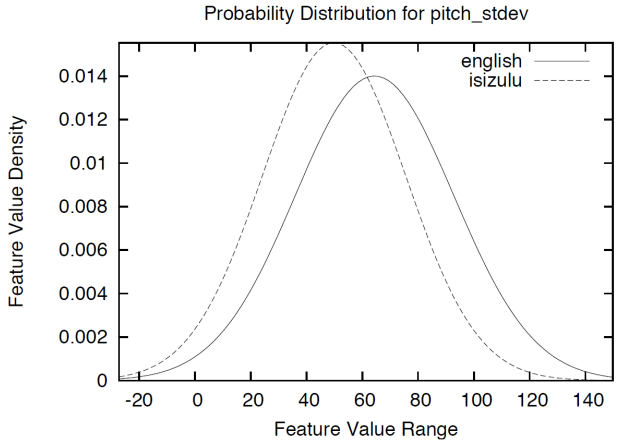

Fig. 4. Value distributions of frequency tremor for South African English and isiZulu of *adult female* speakers. The separation is not as clear as in Fig. 3, but it still shows a considerable difference.

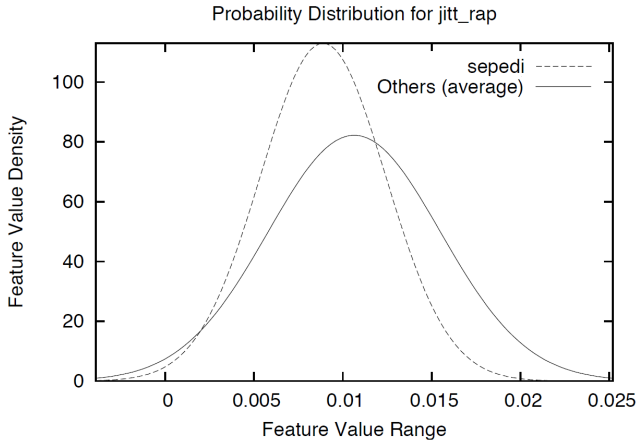

Fig. 5. Jitter (here: relative average perturbation, RAP) as an example of a criterium where one language (Sepedi) has an average that is rather distant from that of all other languages. The statistics contains only speakers of the *adult male* class.

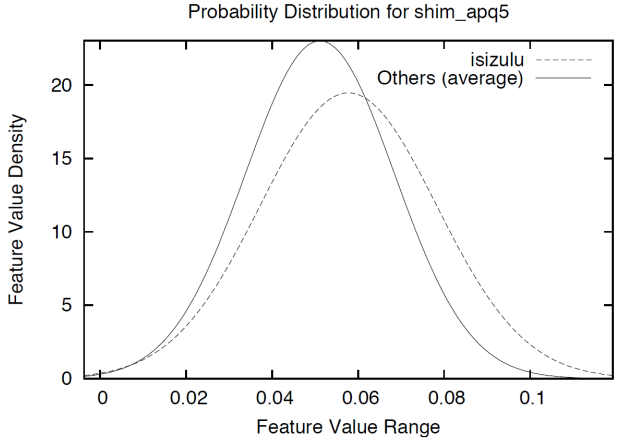

Fig. 6. For *adult female* Zulu speakers, the value range of shimmer (here: the amplitude perturbation quotient for 5-point periods, APQ5) is characteristically higher than for the rest of the languages.

IV. Experimental approach: regression analysis

To investigate the feasibility of age prediction with the long-term features described in Section III, and to compare the different language families in this respect, we developed simple least-squares linear regressors using training data from each of the six languages listed in Table IV. (These languages were selected since they span a variety of the (sub-)families found in the Lwazi corpus, and reliable meta-data was available for their speakers.) Two measures of predictive accuracy (mean absolute error and correlation coefficient) were then computed, employing the models trained on each language separately on the test data from all languages. In particular, the following steps were carried out:

- Each of the six sub-corpora were divided into training and test sets; the ratio of training to test data was approximately 80:20, with no speaker overlap between these sets. The feature vectors listed in Table III were calculated for each utterance in the training and test sets of all languages.
- Each training set was scaled separately so that each feature has zero mean and unit variance; for each language α the regression vector w_α was then calculated as

$$w_\alpha = (X_\alpha^t X_\alpha)^{-1} X_\alpha^t t_\alpha, \tag{1}$$

  where X_α is the matrix formed by stacking all the scaled feature vectors (each extended with a "bias" term of 1) together and t_α is a vector consisting of all the true ages corresponding to the feature vectors in X_α.
- The ages of the speakers of all utterances in the test sets were estimated as

$$y_{\alpha\beta i} = x_{\beta i}^t w_\alpha, \tag{2}$$

  where $x_{\beta i}$ is the extended feature vector for speaker i from language β. For each pair (α, β), we calculated the average of the absolute difference between the estimated and actual ages for all utterances in the test set, as well as the Pearson correlation coefficients between the estimated and actual ages.

TABLE III
Feature Vector Used in the Regression Experiment; Full Definitions of these Features are Available in [10]

#	Feature	#	Feature	#	Feature	#	Feature
1	pitch_min	8	intens_mean	12	jit_l	16	shim_l
2	pitch_max	9	intens_min	13	jit_la	17	shim_ldb
3	pitch_quant	10	intens_max	14	jit_ppq	18	shim_apq3
4	pitch_mean	11	intens_stddev	15	jit_rap	19	shim_apq5
5	pitch_stddev					20	shim_apq11
6	pitch_mas						
7	pitch_swoj						

V. Results

Figures 7 and 8 show the mean prediction errors and correlation coefficients resulting from our linear regression, respectively. We see that the highest accuracy by both measures is generally achieved when training and test sets are drawn from the same language, suggesting that the age factors

TABLE IV
DATA USED FOR REGRESSION TRAINING AND TESTING

Language	# training speakers / utterances	# test speakers / utterances	Mean age	Std dev age
Afrikaans	159 / 4767	40 / 1193	34.7	14.3
English	155 / 4623	37 / 1105	37.7	15.8
isiZulu	149 / 4349	38 / 1122	35.4	14.1
isiXhosa	131 / 3865	34 / 1007	36.8	10.3
Sesotho	153 / 4576	36 / 1049	34.3	12.7
Setswana	152 / 4481	39 / 1149	36.0	13.6

are expressed differently in the different languages. When training and test languages agree, the correlation coefficients range between approximately 0.2 and 0.36, suggesting that this is a challenging task; the corresponding mean values of the absolute errors range between 7.7 and 12.8 years. (As a basis for comparison, when we apply these same methods to a previously-used corpus of German utterances, we obtain correlation coefficients and mean absolute errors of 0.38 and 17.2, respectively. The range of ages in that corpus is substantially larger than in Lwazi, which explains both the higher correlation coefficient achieved – since it is easier to predict more extreme ages – and the larger mean absolute error.)

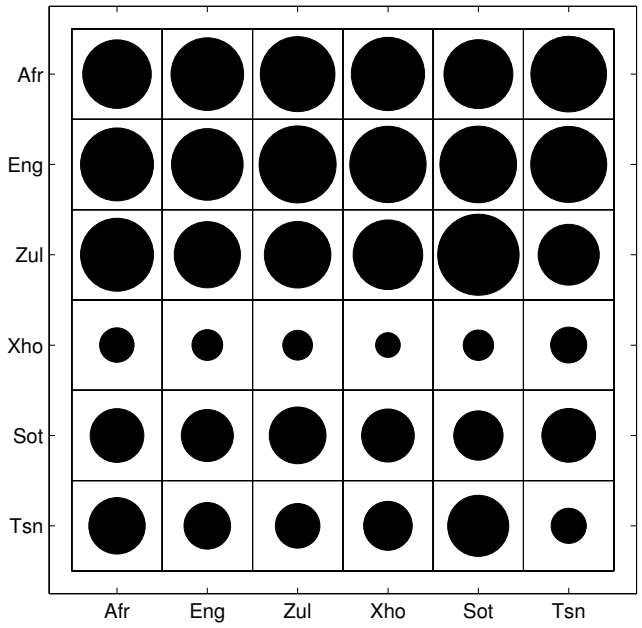

Fig. 7. Mean absolute prediction errors when linear regressor from one source language is applied to all six target languages. The rows correspond to error values for the same target language, and the columns correspond to the language used for training.

When comparing the cross-language predictors, it is interesting to note that the language families are apparently not particularly relevant to age prediction. Thus, the predictor for English ages with the largest correlation coefficient is derived from Sesotho data, and the isiZulu and Setswana predictors are quite accurate when applied to data from the other language in this pair. In contrast, the Sesotho and Setswana regressors

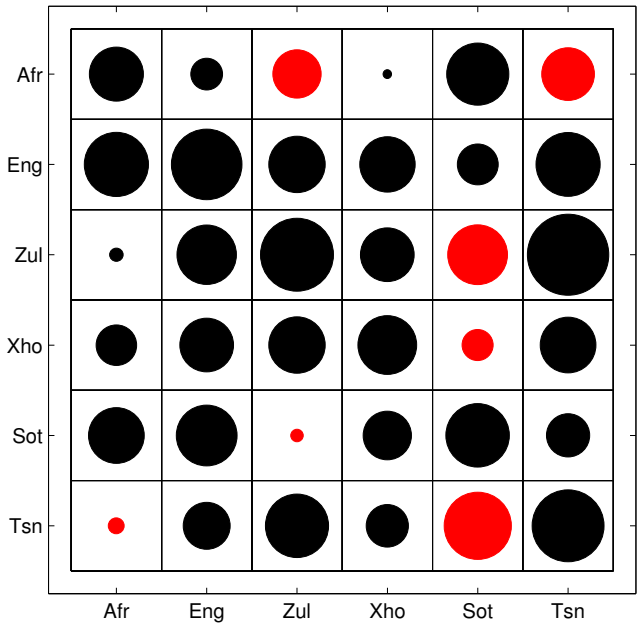

Fig. 8. Correlation coefficients for the same conditions as those in Fig.7. Negative values are indicated by red areas.

do not perform well when applied to test data from the other language in this closely-related pair of languages.

Figure 9 shows the regression weights w_α calculated for all languages. Since all features were normalized to have the same mean and variance, these weights are directly comparable. Many features show significant variation across the different languages. The most consistently important value corresponds to feature 13, which is a long-term average of the jitter in pitch frequency; however, even that feature contributes little to the isiZulu regressor. Feature 10 (related to the minimum intensity within an utterance) has a large negative contribution for English, but contributes somewhat positively to the age regressors for isiZulu and isiXhosa.

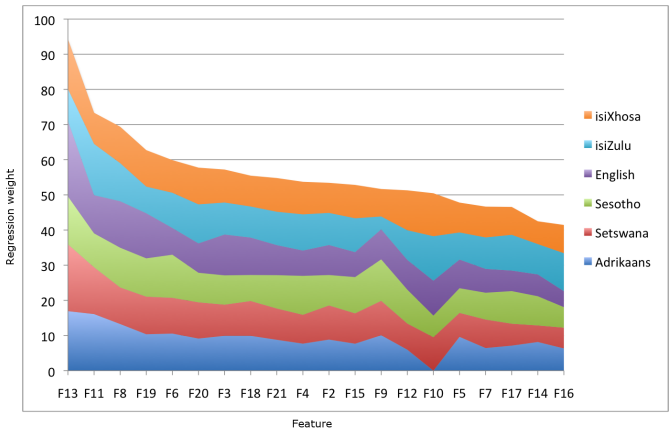

Fig. 9. Regression weights calculated on training data from six different languages

VI. Conclusion

The results from the distribution analysis show that the various languages do not behave in a consistent fashion with respect to age changes. Thus, even this basic para-linguistic information source seems to have significant language-specific (or culture-specific) aspects. The differences between the different features in this respect suggest that in the end, a single feature set may not provide the optimal performance for all languages. Further investigation along these lines would be most interesting, and should include studies on MFCCs as an alternative set of features.

The cross-language comparisons of the age regressors show that the best predictions (in terms of both measures employed in our study) are obtained when training and test data are drawn from the same language. This confirms that the age predictors, in terms of the features employed here, are somewhat language dependent – a conclusion that is further strengthened by the fact that the regression vectors have significantly different shapes for the different languages. However, the predictors are not particularly accurate when applied to test data within the same language family. This observation may indicate that there are other relevant variables – possibly cultural or socio-economic – which play an important role in the observed inter-language differences.

References

[1] M. Boden, *Mind as Machine: A History of Cognitive Science*. New York, NY: Oxford Univ. Press, 2008, ch. 9.

[2] J. R. Bellegarda, "Language-independent speaker classification over a far-field microphone," in *Speaker Classification II: Selected Projects*, C. Müller, Ed. Berlin: Springer-Verlag, 2007, pp. 104–115.

[3] N. T. Kleynhans and E. Barnard, "Language dependence in multilingual speaker verification," in *Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa*, Langebaan, South Africa, Nov. 2005, pp. 117–122.

[4] M. Shami and W. Verhelst, "Automatic classification of expressiveness in speech: A multi-corpus study," in *Speaker Classification II: Selected Projects*, C. Müller, Ed. Berlin: Springer-Verlag, 2007, pp. 43–56.

[5] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Nöth, "Age and Gender Recognition for Telephone Applications Based on GMM Supervectors and Support Vector Machines," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, 2008.

[6] Meraka-Institute, "Lwazi ASR corpus," 2009, online: http://www.meraka.org.za/lwazi.

[7] E. Barnard, M. Davel, and C. van Heerden, "ASR corpus design for resource-scarce languages," 2009, accepted for publication, Interspeech.

[8] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, Aug. 2001.

[9] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus, NIST order number PB91-100354," February 1993.

[10] 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.

[11] P. Boersma, *Praat, a system for doing phonetics by computer*. Amsterdam: Glott International, 2001.

[12] C. Müller and F. Burkhardt, "Combining Short-term Cepstral and Long-term Prosodic Features for Automatic Recognition of Speaker Age," in *Proceedings of the Interspeech 2007*, Antwerp, Belgium, 2007.