Modeling the Speaker-specific F0 Changes Caused by Raised Vocal Effort

C. Harwardt[1]
[1] *Research Establishment for Applied Science (FGAN), Germany, Email: harwardt@fgan.de*

Introduction

F0 measurements and associated features are widely used in automatic and forensic speaker recognition ([1],[2]). One measurement often used is the F0 mean value [3]. This measurement enables a reasonable distinction between target and non-target speakers for audio signals with similar vocal effort. Comparing two signals with different degrees of vocal effort can cause great degradations of recognition results in automatic speaker recognition [4] and requires special awareness when using F0 measurements in forensic speaker identification. Therefore the problem of raising vocal effort and its influence on F0 measurements is investigated continuously in forensic phonetics (e.g. [5] and [6]).
The specific objective of this study is to present a solution to estimate a speaker's F0 variation caused by a specific raise of vocal effort with the aim to improve F0-based speaker recognition tasks. The basic hypothesis of this study is that the ratio between power and F0 mean values in normal speech is closely related to the same ratio for speech produced with increased vocal effort.
The research areas of inter- and intra-speaker F0 variability, F0 variation caused by increasing vocal effort, and modeling these within and across speaker variations are the preconditions for our assumptions. A good overview over past F0 studies including their F0 mean value and the average F0 variation is given in [7]. The special case of variability caused by changing vocal effort is explicitly described in [5]. The most important statement of [5] for this study is that increasing the vocal effort led to a raise of F0 mean values for all 100 analyzed speakers. Furthermore the degree of upraising F0 is speaker-dependent and cannot be modeled globally. This statement is one of the preconditions for the speaker-specific F0-power ratio proposed in this paper.

Modeling intra-speaker-specific F0 changes

One method for modeling within and across speaker F0 variation caused by increased vocal effort is presented in [8]. The presented estimation techniques are proven by tests with 15 Dutch speakers (7 male, 8 female) reading various different sentence types repeating them several times. These sentence types included utterances of different lengths and statements with explicit contrast. The main difference to this study is, that they use read utterances with identic texts as basis for normal and loud speech, whereas we use different utterances in spontaneous speech. Furthermore they predict target F0 values instead of F0 mean values. [8] proposes four different estimation methods for prediction of target F0 values. Their most successful method is a linear raising function including two speaker-specific factors. The first factor is the degree by which the individual speaker raises his F0 range from normal to raised mode. The second factor models the shifting of the F0 range. Reducing the number of these factors or rather modeling cross-speaker relationships instead of exclusively using speaker-specific parameters leads to a degradation of the estimation. Therefore we can conclude that the speaker-specific difference in raising F0 when raising vocal effort is an important feature for describing or predicting such F0 changes.

The data

The data used in this study derives from the Pool 2010 corpus [5]. The Pool 2010 corpus contains audio data from 100 male native speakers of German. There are four audio recordings available for each speaker. The different modes recorded cover read and spontaneous speech, each combined with the two modes normal speech and speech with increased vocal effort. The increase of vocal effort was induced by exposing 80 dB white noise to the speakers via headphones. For this study we used the spontaneous speech with normal and increased vocal effort from 100 speakers transmitted via GSM.
As measurement of vocal effort we extracted the power of each audio signal. We used this kind of measurement because recording level and distance to the microphone were kept constant for all recordings, so that this measuring procedure leads to reasonable results. The pitch extraction was done with the ESPS method of the Snack toolkit [9]. The output of this method contains the pitch as well as a probability of voicing for each sample. This probability was used to dismiss the power and pitch measurements of the non-voiced parts of the audio files.

Estimating speaker-specific changes in F0 mean

As already shown in [8], prediction of F0 changes caused by increasing vocal effort includes speaker-specific parameters, which, if they are eliminated, cause a degradation of prediction results. Thus we conclude, that sufficient modeling of the speaker-specific F0-power ratio leads to a reasonable F0 prediction, which can be used in automatic as well as in forensic speaker recognition tasks. This F0-power ratio should result in a low deviation between estimation and actual F0 mean value for target speaker trials. For non-target speaker trials this deviation should be larger. Otherwise, if target

and non-target speaker trials lead to similar deviation values, the ratio would be a global, not a speaker-specific parameter. Thus, a sufficient distinction between target and non-target trials by means of deviation values could not be guaranteed.

Method

Considering a speaker recognition task, we will below refer to the signal with normal speech as training signal and to the signal with raised vocal effort as test signal. The test signal might (in case of the target trials) or might not (in case of the non-target trials) be spoken by the same speaker as the training signal. This results in 100 target trials and 9900 non-target trials for the 100 speakers of the Pool 2010 corpus. The focus of this paper lies on the F0 mean prediction of the target trials to verify the proposed estimation method. The non-target trials are presented to prove the speaker-specific characteristics of the used method.

To generate the speaker-specific F0-power ratio we divide the data of the training signal into two equal parts (tr1; tr2). We then calculate the mean values of F0 and power for the voiced parts of the whole training and test signal as well as for the divided signals. The F0 mean value of the test signal is calculated as a reference value for the estimation. It will not be used in the estimation procedure. The division of the training signal simulates two different signals of one speaker that can be used to make an estimation of the target speaker's F0 changes. To get this estimation we divide the F0 and power mean values of tr1 by the mean values of tr2. To get the ratio for the training signal R_{tr}, we divide the F0 ratio derived by the divided training signals by the power ratio also derived by the divided training signal and calculate the absolute value. The power ratio R_p is calculated by dividing the mean power value of the whole train signal by the mean power value of the test signal. The speaker-specific F0-power ratio of the test signal R_{fp} is defined as follows:

$$R_{fp} = \mid \frac{R_{tr}}{R_p} \mid \qquad (1)$$

R_{fp} is used as a speaker-specific raising factor to predict the F0 mean value that the speaker of the training signal would produce if he would produce speech with the power mean value measured in the test signal. This implies that for the prediction of the F0 mean value for the test signal, we multiply the F0 mean value of the complete training signal by R_{fp}.

Results

We tested our estimation method on 100 male speakers of the Pool 2010 corpus.

We calculated the deviation of the predicted F0 mean from the actual mean. This analysis resulted in a mean deviation for target trials of 10.5%. For non-target trials the mean deviation value is 17,6%. These mean values show that the deviation for target trials is in average 41.5% lower in comparison with the deviation for non-target trials.

The standard deviation of the target trials (12.1%) is, like the mean value, smaller than the standard deviation for non-target trials (15.6%).

A good summary of the overall prediction performance for target and non-target trials is shown in figure 1.

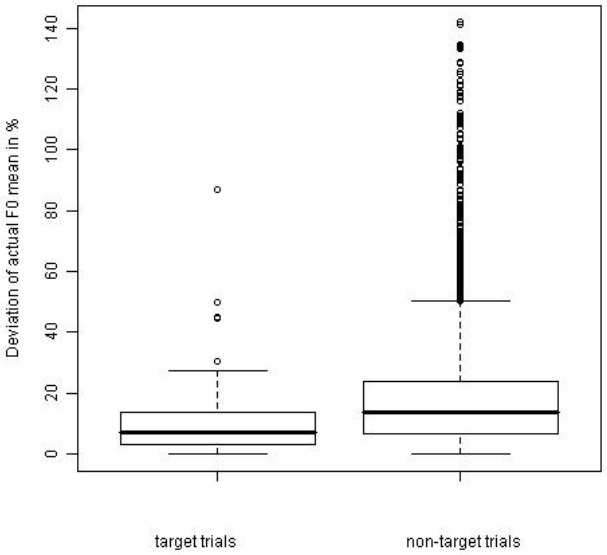

Figure 1: Boxplot of target and non-target trials

Figure 1 presents the deviation between predicted F0 mean and actual F0 mean in percentage. The figure illustrates that for the target trials 50 percent of the deviations lie between 3.3 and 13.6% with a median of 7%.

For the non-target trials 50 percent of the deviations are ranging between 6.5 and 24%. The median is 13.9%.

The most important result shown in figure 1 is that there are just four outliers for the target trials but a great bulk of outliers for the non-target trials. The four outliers of the target trials show that for those four speakers no adequate approximation can be reached with the given prediction method. The bulk of outliers of the non-target trials shows that the non-target speaker trials often lead to F0 mean values that are more than two times bigger than the actual F0 mean value. This great difference makes it easy to classify such trials as non-target.

From all the given facts we can conclude that the proposed prediction method is a good first approach for the estimation of F0 mean in high-effort speech. The deviations of target trials are in average much lower than those for non-target trials and therefore support the thesis of speaker-specific instead of global character of the prediction method. As future work, a more detailed prediction method could help to refine the differentiation between target and non-target trials in reducing the deviation for target trials and expanding it for non-target trials.

Conclusion and future work

In this paper we pointed out the problem of F0 mean raise caused by increased vocal effort. We presented

an estimation technique to predict F0 mean for a given target speaker. To prove this estimation technique, we presented results of using this technique on audio recordings of 100 male speakers. Doing target and non-target tests we proved the speaker-specific and non-global character of the prediction method. On the whole, the presented method can be seen as preliminary work which should be refined further to improve the differences between target and non-target trials.
Further work to be done is the investigation of measuring procedures for sufficient classification of the degree of vocal effort in non-laboratory speech. Another objective of this study which we haven't investigated yet is to be able to make predictions independent of which kind of audio signal (loud or normal) is to be used as training data. The user should be able to make predictions both ways, from normal to loud speech and vice versa. Additionally, the approach should be tested on signals with different lengths of training and test data.

References

[1] Elisabeth Shriberg. (2007). *Higher-Level Features in Speaker Recognition.* in: Speaker Classification I - Fundamentals, Features, and Methods, eds: Christian Müller.

[2] Philip Rose. (2002). *Forensic Speaker Identification.* London & New York: Taylor & Francis.

[3] Angelika Braun. (1992). *Zur Bedeutung des Merkmals "mittlere Sprechstimmlage" in der forensischen Sprechererkennung.* in: Phonetik und Dialektologie, eds: Dingeldein.

[4] Timo Becker. (2008). *The influence of intra-speaker variability in automatic speaker verification using F0 features.* in IAFPA 2008.

[5] Michael Jessen, Olaf Köster, and Stefan Gfroerer. (2005). *Influence of vocal effort on average and variability of fundamental frequency.* International Journal of Speech, Language and the Law, **12**, 174-213.

[6] P. Gramming, J. Sundberg, S. Ternström, R. Leanderson, and W. H. Perkins. (1987). *Relationship between changes in voice pitch and loudness.* in STL-QPSR, vol. **28**, pp. 39–55.

[7] Hartmut Traunmüller and Anders Eriksson. (1995). *The frequency range of the voice fundamental in the speech of male and female adults.* http://www.ling.su.

[8] Elizabeth Shriberg, D. Robert Ladd, Jacques Terken, and Andreas Stolcke. (1996). *Modeling Pitch Range Variation within and across Speakers: Predicting F0 Targets when "Speaking Up".* Proceedings of the fourth International Conference on Speech and Language Processing.

[9] Kare Sjölander and Jonas Beskow. (2000). *WAVESURFER AN OPEN SOURCE SPEECH TOOL.* Proceedings of ICSLP 2000, 6th Intl Conf on Spoken Language Processing. B. Yuan, T. Huang, and X. Tang, Eds., Beijing, pp. 464–467.