Proceedings of DiSS-LPSS Joint Workshop 2010,
25-26 September 2010, Tokyo, Japan, pp. 35-38.

Investigating the COG Ratio as Feature for Speaker Verification on High-Effort Speech

Corinna Harwardt

Fraunhofer FKIE, Command and Control Information Systems, Germany
corinna.harwardt@fkie.fraunhofer.de

Abstract

Vocal effort mismatch in training and test data leads to immense degradations of speaker recognition systems. The changes on the acoustics of a speech signal induced by raised vocal effort are complex and despite several studies from various authors not completely known yet.

Instead of just gaining knowledge about these differences for automatic speaker recognition it is rather an essential to discover features that remain relatively stable in changing vocal effort conditions and contain speaker specific information. In this study we investigate the center of gravity (COG) ratio for high and mid frequency bands as feature for speaker recognition. We find that vocal effort mismatch leads to an equal error rate (EER) more than six times higher for a standard MFCC-based GMM-UBM system. For the COG ratio we observe a much smaller degradation of around 25%.

When adapting the UBM with additional high-effort speech data the EER of the COG ratio gets even better for the mismatch condition than for the matching task. Combining MFCC and the COG ratio leads to best results with an overall improvement of 16% compared to the standard MFCC-based system.

Index Terms: vocal effort, speaker recognition, center of gravity ratio

1. Introduction

Automatic speaker verification already yields good results for several tasks on spontaneous speech. However, a speaker produces many variations in spontaneous speech which can't be captured adequately with standard speaker recognition systems. Such variations might be for example disfluencies, emotions, influence of alcohol or drugs and others. The fact that often more than one of these variations occurs in spontaneous speech makes the investigation of spontaneous speech so challenging. In this paper we take into account just one variation: the change of vocal effort in spontaneous speech.

Vocal effort is the quantity a speaker raises his voice to adopt the loudness of his speech to the actual communication situation. A change of the communication situation, which induces an adjustment of vocal effort, might be for example a variation of the communication distance between the communication partners, hearing impairment, stress and other emotions or background noise. In this study we focus on high-effort speech induced by background noise, the so called Lombard speech [1]. We try to find a robust feature for a speaker verification scenario when the vocal effort does not match in training and test data. We consider normal-effort speech as training data and high-effort speech as test data. This scenario might be interesting for forensic case work, because the offenders' speech sample is often spoken with high vocal effort (e.g. when the offender makes the offence call from his mobile phone on a crowded place), whereas the suspect recordings are typically produced with normal vocal effort.

To be able to develop sufficient features for speaker verification with high-effort speech one should know which changes are induced by high vocal effort to the production and perception of speech. These differences are described in several studies, but due to the complexity and different presuppositions the results are not always consistent. Despite this we try to summarize the changes of the acoustics with focus on formants, pitch and spectral characteristics.

Changing vocal effort leads to several modifications in the acoustics of spontaneous speech. One major change is the higher F0 in high-effort speech [2]. Furthermore the formants are influenced by vocal effort, but compared to F0 these changes are not so clearly definable. The first formant increases in high-effort speech [3, 4]. For the second formant some authors don't notice any significant change [3], whereas others discover individual changes per phoneme [5]. Similarly some authors observe individual changes for F3 [6], other do not find significant changes [3, 4] and some report a shift of F3 to around 2600 Hz [7].

The distribution of energy in the speech spectrum has been part of different studies, too. Concerning the energy most authors go confirm with each other. They describe a tendency to shift the energy from the lower to the mid and higher frequencies in high-effort speech (e.g. [3]). This energy migration leads to changes of spectral features. Spectral tilt, which can be described as the slope of the spectral distribution, can be used to distinguish between vocal efforts [8]. Other studies focus on variations of the energy ratio, spectral balance or other spectral features. The COG which represents the weighted mean of the spectrum is further described in [9, 3]. The spectral center of gravity (COG) calculated over the whole spectrum is influenced by vocal effort too and might be additionally an indicator for stressed speech.

The modifications induced by change of vocal effort affect the traditional cepstral based features in speaker recognition [10]. By this study we want to address this problem and evaluate whether the COG ratio is suitable as feature for speaker verification with vocal effort mismatch.

We first describe the COG ratio as feature for speaker verification and our motivation to use the COG ratio. Then we describe the experiments including the corpora we used for our tests, the setup and the results. The results are divided into three subsections according to the different feature variations. Finally we draw some conclusions and give future perspectives.

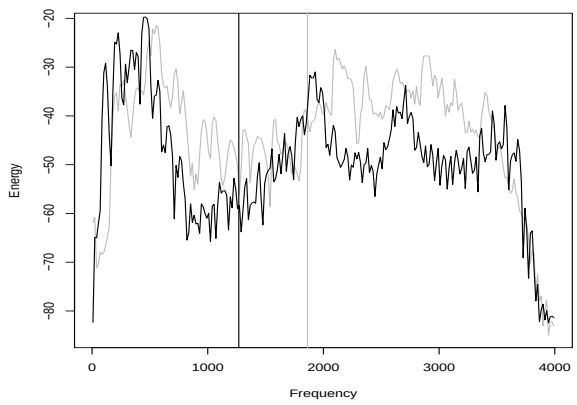

Figure 1: *Spectrum of the vowel [ə] pronounced with normal (black) and high (grey) vocal effort by the same speaker. The COG is marked as a vertical line for each spectrum.*

2. Center of Gravity Ratio

The spectral COG of a speech signal represents the weighted mean frequency of the spectrum. As in [9] we calculate the COG by

$$COG = \sum f_i * E_i / \sum E_i \quad (1)$$

with f_i representing the frequency and E_i standing for the spectral power as a function of the frequency. For energy migrations from low to high frequencies, as observed for high-effort speech, [3] the COG changes. Figure 1 illustrates the changes of COG in the speech spectrum of the vowel [ə].

As mentioned earlier we don't want to find features that change with vocal effort. Therefore we do not calculate the COG on the whole spectrum. As we know that the high and mid frequency components of the spectrum both get enriched in high-effort speech, we divide the ratio of the COG of high frequencies by the COG of mid frequencies. The choice of the frequency bands is motivated by the formant locations and by the fact that most speaker recognition applications must use band limited telephone speech. Hence the high frequency band contains frequencies from 2200 - 3000 Hz whereas the mid frequency band covers the frequency range from 800 to 2200 Hz. Figure 1 presents an example for the raise of the energy in the two frequency bands for the vowel [ə].

For the use of the COG ratio as feature in speaker verification systems we calculate the delta and delta delta features for each COG ratio over a five frame context. The resulting three dimensional vector is used as feature vector for the verification task.

3. Experiments

3.1. The Corpora

The data used in this study derives from the Pool 2010 corpus [2]. The Pool 2010 corpus contains audio data from 105 male native speakers of German. For each speaker four audio recordings are available. The different modes recorded cover read and spontaneous speech, each combined with the two modes normal speech and speech with increased vocal effort. Increase of vocal effort was induced by exposing 80 dB white noise to the speakers via headphones. For this study we used spontaneous speech with normal and high vocal effort transmitted via GSM. The data has been divided into a development set of 55 speakers and a test set of 50 further speakers. For training we used one recording per speaker containing about 50 seconds of normal-effort speech. For the tests with high-effort speech we had two recordings, each containing again around 50 seconds of speech. To get an impression of the systems performance loss due to vocal effort changes we needed another test set with normal-effort speech. This test set contained one recording per speaker, again with 50 seconds speech.

For the UBM (universal background model) we used additional data from the Kiel corpus [11]. The Kiel corpus consists of normal-effort speech. Both corpora contain German speech.

3.2. Experimental Setup

To test the COG ratio as feature for speaker verification we used a statistical framework based on gaussian mixture models (GMM) as proposed by [12]. First we trained an UBM that represents the speech and language from the speaker population under consideration, in this case German male speakers. To train this UBM we used the male speakers from the Kiel corpus. Next we adapted the speaker models from the well trained UBM with normal-effort speech from each of the 50 speakers from the Pool 2010 corpus. Lastly we ran the tests on the doubtful data. These doubtful data were either normal- or high-effort speech samples from the Pool 2010 corpus.

In this study we varied the features which are needed to train the models and to run the verification. The tests we ran are:

- Tests with standard MFCC features. The MFCC feature vector contained 15 MFCC and their first and second order derivate.
- Test with the COG Ratio. When using the COG ratio as feature we calculated a three dimensional feature vector with the COG ratio and the first and second order derivate.
- We extended our UBM training data by additional high-effort speech and adapted the existing UBM with this data. The adaptation data consisted of the development set from the Pool 2010 corpus.
- Additionally we fused the results from both systems to see whether we can improve the overall performance.

The results from these tests are described in the next sections.

3.3. Results

In the next subsections we will illustrate the experimental results with different features. First of all we describe the results with the standard MFCC features followed by the results with the newly proposed COG ratio and the combination of both features. The results of the adaptation are not listed separately. They are included in the description of the single features.

3.3.1. MFCC

In this subsection we describe MFCC as features for speaker verification with changing vocal effort conditions. We included the MFCC into a GMM-UBM system with 1024 mixture components. For the training procedure we used normal-effort audio data. As depicted in table 1 we reach an EER of 0.57% for normal-effort speech in training and test data. Usage of high-

Table 1: *EERs for the MFCC-based system.*

vocal effort	EER
normal-effort	0.57%
high-effort	4%
high-effort (adapted)	4%

effort speech as test samples raises the EER to 4%. Doing a MAP (maximum a posteriori) adaptation of the UBM with additional high-effort speech from the Pool 2010 corpus before training the speaker models does not improve the performance of the system (see table 1).

3.3.2. COG Ratio

Usage of the COG ratio is motivated by the fact that mid and high frequencies (as defined in section 2) in the speech spectrum get enriched if a speaker raises his vocal effort. Our underlying hypothesis for using this feature is that the ratio of the COG of these frequency regions seems to be relatively stable and additionally the COG ratio seems to have a great inter-speaker variability. If this thesis is correct the COG ratio would be a good feature for this speaker verification scenario. In table 2 we plotted the results of our tests. The standard GMM-UBM

Table 2: *EERs for the COG ratio based system. The number of mixture components is plotted in the rows and the kind of test data in the columns. For training we used always normal-effort audio data.*

	normal-effort	high-effort	high-effort (adapted)
64	24%	28%	23%
128	22.49%	28%	22%
256	24.65%	28%	23%
512	22.82%	29%	23%
1024	22%	29%	24.08%

systems operating with MFCC use 1024 or 2048 mixture components. Of course the feature vectors used in these systems are much larger than the features used for COG ratio, which have just three elements. Therefore we tried different numbers of mixture components and found a number of 128 mixtures to give best results. As shown in table 2 the system operating with 1024 components yields better results for normal-effort speech, but it is worse for high-effort speech. Additionally it has the disadvantage of a longer training procedure.

When comparing the results of normal and high-effort test samples we see a degradation of the performance for all the system variants. However, the degradation is not as great as for the MFCC-based system which has a more than six times higher EER for high-effort test speech. Hence we conclude that this feature is relatively robust against vocal effort mismatch. When we compare the EERs to those of the MFCC-based system we observe a much higher EER for the COG ratio. One reason for this might be the number of elements of feature vectors. To achieve better results with the COG ratio we should use them in combination with other features. One solution might be the combination of MFCC and COG ratio (see next subsection 3.3.3).

We observed the best results when adapting the UBM with additional high-effort speech data before speaker model training. This adaptation has the advantage that we don't need high-effort speech per speaker as proposed in [13]. We can rather use an extended UBM to get better results. For 128 mixtures the results on high-effort speech do even get better than the results on normal-effort test speech.

3.3.3. Fusion

To test whether the COG ratio can improve the performance of the standard MFCC-based system we fused the scores of both systems. For these merged scores additional audio data for training is needed, which should be as similar as possible to the data used later on. On the other hand the data used for other development purposes (e.g. the MAP adaptation of the UBM) should not be reused to train the fusion of scores. In the context of this work only the results of the high-effort test samples are regarded as important. Therefore we used the normal-effort speech samples as training data for merging the high-effort scores. We utilized the FoCal toolkit[1] for the fusion. The process is a linear fusion as described by [14].

The EERs of the single, as well as the fused features are presented in table 3. The MFCC system always operates with

Table 3: *Comparison of the EERs of the different systems on high-effort speech.*

Features	EER
MFCC	4%
COG Ratio	28%
COG Ratio (adapted)	22%
MFCC + COG ratio	3.35%
MFCC + COG ratio (adapted)	4%

1024 mixtures, whereas the COG ratio is used in conjunction with a 128 mixture model. As you can see in table 3 the fusion outperforms the system based on the COG ratio as well as the standard MFCC system. During the fusion process the scores get simultaneously calibrated by the FoCal toolkit. To ensure that the improvement doesn't originate in the previous calibration, we calibrated the scores of the MFCC-based system. Because of quantitative lack of training data (the same as for the fusion) it was not possible to perform a sufficient calibration.

Using scores from the adapted COG ratio system does not lead to an improvement over the MFCC alone. To provide a better performance overview, figure 2 plots the results as DET curves (detection error trade off). The systems operating on the COG ratio only are worse than those using the MFCC. The best performance is achieved by the system which makes use of both the MFCC and the COG ratio.

4. Conclusion

We evaluated the COG ratio as feature for speaker verification on high-effort speech. For this we used a GMM-UBM speaker verification framework with MFCC features as baseline.

Next we included the COG ratio and the first and second order derivate as features in this framework. We observed that compared to the MFCC features the COG ratio was relatively stable to changes in vocal effort.

We were able to further improve our results by adapting the UBM with additional high-effort speech data. By this MAP adaptation we gained better results in high-effort speech compared to the baseline test with normal-effort test samples. Hence

[1] http://www.dsp.sun.ac.za/ nbrummer/focal/

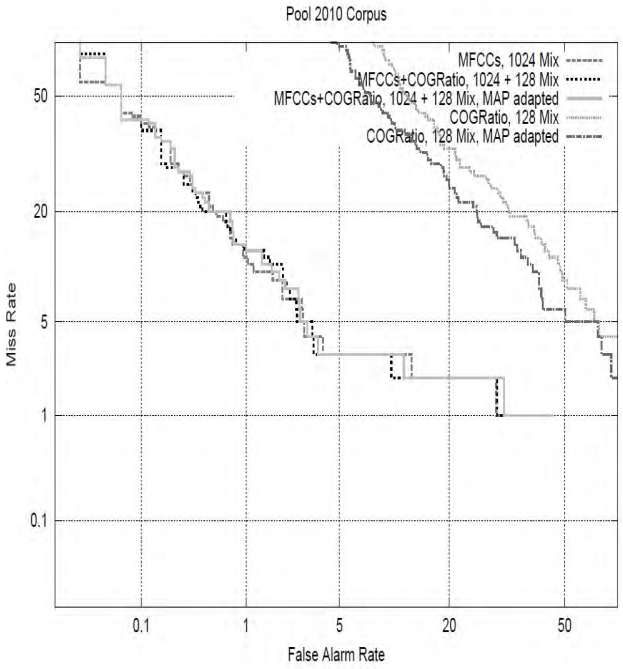

Figure 2: *DET curve for MFCC, COG Ratio, the fusion of both features and the adapted COG ratio.*

we conclude that the degradation due to vocal effort mismatch for COG ratio can be compensated by MAP adaptation of the UBM. The great advantage is that the user doesn't need additional high-effort speech for each enrolled speaker. He just needs additional data from some other speakers. For the MFCC the MAP adaptation of the UBM wasn't successful.

Comparing the EERs of MFCC and the COG ratio we observe that the performance of the COG ratio is much lower although it is stable to vocal effort changes. One reason might be that the MFCC vector consists of 45 components whereas the COG ratio vector has just three elements. In future work we will try to find further features which are stable to vocal effort changes and have a high inter- and low intra-speaker variability. These features could be used to extend the feature vector.

One other possibility is to fuse the scores generated with the COG ratio with other systems' results. We presented a linear fusion of MFCC and COG ratio in this paper. Linear fusion of both systems scores' yielded best results compared to the single systems. We could reach an improvement of around 16% compared to the standard MFCC-based system.

As future work tests with female speakers should be taken into account. Especially tests with both genders mixed in the training and test data are challenging with high-effort speech because male speakers are often identified as female person due to the raised F0. We need to check whether the COG ratio is robust against such gender-based false alarms.

5. References

[1] Lombard, Étienne. "Le signe de l'lvation de la voix", Ann. Mal. Oreil. Larynx 37, 1911, S. 101-119.

[2] Jessen, M. and Köster, O. and Gfroerer, S., "Influence of vocal effort on average and variability of fundamental frequency", International Journal of Speech, Language and the Law, 12:174–213, 2005.

[3] Liénard, J-S. and Di Benedetto, M-G., "Effect of vocal effort on spectral properties of vowels", J. Acoust. Soc. Am., 106(1): 411–422, 1999.

[4] Schulman, R., "Articulatory Targeting and Perceptual Constancy of Loud Speech", Phonetic experimental research at the Institute of Linguistics, University of Stockholm, 1985.

[5] Bond, Z. S. and Moore, T. J. and Gable, B., "Acoustic-phonetic characteristics of speech produced in noise and while wearing an oxygen mask", J. Acoustic Soc. Am., 85 (2): 907-912, 1989.

[6] Stanton, B. J. and Jamieson, L. H. and Allen, G. D., "Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions", ICASSP 88, New York City, 1988.

[7] Geumann, A., "Vocal intensity: acoustic and articulatory correlates", Conference on Motor Control, Nijmegen, 2001.

[8] Zhang, C. and Hansen, J.H.L., "Analysis and Classification of Speech Mode: Whispered through Shouted", Proc. Interspeech, pp.2289-2292, 2007.

[9] van Son, R. J. J. H. and Pols, L. C. W., "An acoustic description of consonant reduction", Speech Communiction, 28: 125–140, 1999.

[10] Shriberg, E. and Graciarena, M. and Bratt, H. and Kathol, A. and Kajarekar, S. and Jameel, H. and Richey, C. and Goodman, F., "Effects of Vocal Effort and Speaking Style on Text-Independent Speaker Verification", Proc. Interspeech, pp. 609612, Brisbane, Australia, 2008.

[11] Kohler, K. J. (editor), "Arbeitsberichte (AIPUK) Nr. 29", Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel, 1995.

[12] Reynolds, D. and Quatieri, T. and Dunn, R., "Speaker Verification Using Adapted Gaussian Mixture Models.", Digital Signal Processing, 10: 19–41, 2002.

[13] Hansen, J.H.L. and Varadarajan, V., "Analysis and Compensation of Lombard Speech Across Noise Type and Levels With Application to In-Set/Out-of-Set Speaker Recognition", IEEE Transactions on Audio, Speech and Language Processing, 17(2):366–378, 2009.

[14] Brummer, N. and du Preez, J., "Application-Independent Evaluation of Speaker Detection", Computer Speech & Language, 20 (2-3): 230–275, 2005.