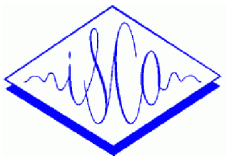

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

The log-Gabor method: speech classification using spectrogram image analysis

Harm Buisman and Eric Postma

Tilburg center for Cognition and Communication, Tilburg University, Tilburg, The Netherlands

harmbuisman@gmail.com, ericpostma@gmail.com

Abstract

We explored the suitability of the log-Gabor method, a speech analysis method inspired by Ezzat e.a. (2007), for automatic classification of personality and likability traits in speech. The core idea underlying the log-Gabor method is to treat the spectrogram as an image of spectro-temporal information. The image is transformed into Gabor energy values using the two-dimensional logarithmic Gabor transform, which is a standard feature extraction method in visual texture analysis. The aggregated energy values are mapped onto classes by means of a support vector machine (SVM). The log-Gabor method performed above baseline on the INTERSPEECH Personality and Likability Sub-Challenges Development sets and comparable to baseline for the Test sets. These results support further investigation of the log-Gabor method as a method for extracting perceptual cues from speech.

Index Terms: spectro-temporal analysis, spectrogram analysis, log Gabor filters, likability classification, personality classification, support vector machines

10.21437/Interspeech.2012-99

1. Introduction

In recent years, several studies investigated the use of the spectrogram (or spectrogram-like transformations) as a basis for extracting perceptual cues from speech [1-4]. The application domains in which these studies were performed include speech intelligibility analysis [2], emotion recognition [3], and automatic speech recognition [4]. The outcomes of these studies show that socially informative perceptual cues can be extracted from the spectro-temporal patterns in the spectrogram.

We study a method for the extraction of such cues that is based on the Gabor transform. A local decomposition in terms of the (logarithmic) Gabor transform has become one of the main transforms for feature extraction in image analysis [5-7]. This transform analyses the visual contours and features by means of spatially localized oriented spatial frequency filters. Since the spectrogram obtained with the short-time Fourier transform (STFT) is two-dimensional, it can be visualized as an image, even though it is not an image in the strict sense [1]. This follows from the fact that important characteristics of speech, such as pitch, rhythm, and attack and decay, reveal themselves in the spectrogram as localized periodic patterns that have orientations and (spatial) frequencies. Pitch is reflected in the harmonics of the fundamental frequency f_0 with a vertical orientation and spatial frequency (spacing of the harmonics). Rhythm is a predominantly temporal pattern, which is expressed in the spectrogram as a horizontal periodic structure. Finally, attack and decay are both temporal and spectral, and as such have an orientation that may have any orientation. The 2D Gabor transform performs local measurements of orientation and spatial frequency and seems therefore suitable for the extraction of common speech characteristics.

This study seeks to determine whether the Gabor transform can be successfully applied to extract perceptual cues to personality and likability. We evaluate the method on the datasets provided as subsets of the INTERSPEECH 2012 Likability and Personality Sub-Challenges [8].

The outline of the remainder of this paper is as follows. In Section 2 we present the log-Gabor method. Then, in Section 3 the experimental setup for evaluating the method on the INTERSPEECH 2012 Challenges is described. Section 4 reports the results on the Sub-Challenges. In Section 5 we discuss the log-Gabor method and its performances and present our conclusion.

2. The log-Gabor method

The log-Gabor method consists of four stages: (1) generation of the spectrogram, (2) application of the logarithmic Gabor transform, (3) feature extraction and (4) classification onto a class. In the following four subsections, each of the stages is discussed in more detail.

2.1 Generation of the spectrogram

In the log-Gabor method the spectrogram is generated by applying the short-time Fourier transform (STFT) to the speech signal. The temporal resolution of the spectrogram is defined by a single parameter Δt, the length of the time segments or time resolution. The spectrogram is created with a Hanning window and an overlap of $\Delta t/2$.

The input for the log-Gabor transform is the logarithm of the spectrogram, as is common in, for instance, automatic speech recognition [4]. For convenience we henceforth refer to the logarithm of the spectrogram simply as the spectrogram.

2.2 Log Gabor transform

The log Gabor transform has its basis in the Gabor function [9], which represents the optimal measurement of location, frequency and orientation. For the log-Gabor method, we employ logarithmic Gabor functions as suggested in [5] and [6]. A multi-resolution filter bank of log Gabor filters provides equal coverage of the Fourier domain with minimal overlap between the filters. The log-Gabor method employs such a filter bank, yielding a filtered "energy image" of the spectrogram for each combination of the n_s scales and n_o orientations. For each filtered image, a specific scale and orientation is retained. Scales can be interpreted as the difference in pixel period of patterns in the image (e.g. high and low pitch harmonics are spaced at a different pixel distance) and orientations can be vertical (e.g. harmonics), horizontal (e.g. rhythm), or oriented (e.g. attack and decay). The energy values of pixels in these images represent the

degree to which the filter-specific spectro-temporal patterns are represented at the pixel location in the spectrogram image. For example, a segment of voiced speech with clearly visible pitch harmonics gives a high response to a vertical filter with a spatial frequency that matches the spacing of the harmonics. Figure 1 shows an example of such a vertical filter, along with an example of a diagonal filter.

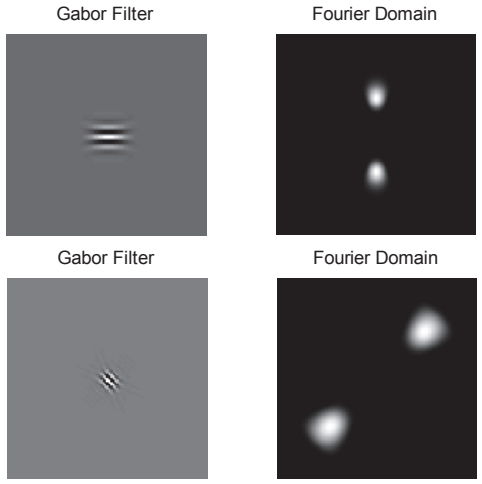

Figure 1: *Left panel: Illustration of two Gabor filters, one vertical (top) and one diagonal (bottom). Right panel: representation of the same filters in the Fourier domain.*

2.3 Feature extraction

Feature extraction comprises (i) extraction of raw features from the energy images generated by the log-Gabor transform, (ii) dimensionality reduction of the raw normalized features, and (iii) normalization of dimensionality-reduced features.

The raw features are extracted from the energy images. They are computed at the level of the whole speech sample by summing the energy values over time. To that end, the energy images are subdivided into n_b equally-sized (on a *log2* scale) frequency bands. Each frequency band (except the highest band) yields a feature by summing over that band. In addition, one feature is computed over the entire energy image (i.e., over all n_b frequency bands). This results in n_b features per energy image. In total, $n_s \cdot n_o \cdot n_b$ raw features are extracted for each speech sample.

The number of raw features can become quite large and there can be considerable correlation between the features, making some of them redundant. Principal component analysis (PCA) is applied to generate reduced features, which may enhance the generalization performance of the classifier. In the log-Gabor method the number of components retained is a parameter (k). As PCA over-represents features with large numerical values, the feature values are normalized, i.e., linearly mapped on the unit interval before applying PCA. Scaling is performed again after PCA, since SVMs are known to perform better on normalized feature sets [10].

2.4 Classification

The reduced and normalized features for each speech sample are classified by means of a support vector machine (SVM). We use SVMs with a radial basis function kernel.

3. Experimental setup

The validation of the log-Gabor method is performed on the *INTERSPEECH 2012 Speaker Trait Challenge*, more specifically on the Personality and Likability Sub-Challenges. These Challenges consists of binary classification tasks in which a collection of speech samples (audio files) have to be classified as *Likable* or *Not Likable* (Likability Sub-Challenge) and as the presence or absence of one of the "big five" personality dimensions *Openness*, *Conscientiousness*, *Extroversion*, *Agreeableness*, and *Neuroticism* (Personality Sub-Challenge), see [8] for a more detailed description.

3.1 Parameters

The experimental evaluation of the log-Gabor method has been performed for a large range of parameter values. Table 1 gives an overview of the parameters that were varied and held constant in order to find the best performing model. We distinguish four types of parameters, one for each stage of the log-Gabor method: (i) spectrogram parameters, (ii) log-Gabor transform parameters, (iii) feature extraction parameters, and (iv) classification parameters. Below, we describe each of these types in detail and refer to Table 1 for the parameter values that were considered.

(i) *Spectrogram parameters.* We varied the time resolution Δt of the spectrogram to determine the most informative temporal resolution, choosing 20 values in the range 0.25-30 ms.

(ii) *log-Gabor transform parameters*. The log-Gabor transform was performed with Kovesi's [11] implementation (with *minWaveLength* = 3, *mult* = 1.35; *sigmaOnf* = 0.8) using fixed values for n_s and n_o. In addition, we examined two mappings of the log Gabor energy images: the (plain) linear mapping and the logarithmic mapping.

(iii) *Feature-extraction parameters*. For principal component analysis, the parameters were the number of principal components k, were $k=0$ represents that no PCA was performed.

(iv) *Classification parameters*. For the classification stage of the log-Gabor method, we used the LIBSVM implementation [10]. The SVM parameters cost c and gamma g were varied in a two-step grid search as suggested by [10]. The values in Table 1 indicate the range on a *log2* scale ($g = 2^G$, $c=2^C$).

Table 1. *Overview of the used parameters*

Stage/parameter(s)	Range (number of values)
(i) Spectrogram	
Δt	0.25, 0.5, 0.75, 1, 1.5, …, 6, 7, 9, 11, 15, 20, 30 ms (20)
(ii) log-Gabor transform	15/8/5 (fixed)
$n_s/n_o/n_b$	linear, logarithmic (2)
Energy mapping	
(iii) Feature extraction	
# of principal components	0, 1, …, 12, 20, 30, 60, 100 (17)
(iv) Classification	
sets	gender separation yes, no (2)
SVM-gamma	G: -7:-1; ±0.1, …, 0.5 (7,10)
SVM-cost	C: -2,..,8; ±0.1, …, 0.5 (11,10)

Finally, to investigate the effects of gender separation, the datasets were either presented as a whole or as separate male and female sets.

3.2 Validation

The INTERSPEECH 2012 Challenge datasets consists of three subsets: training set, development set, and test set. Of these subsets, the training and development sets are labeled. The test set is unlabeled. We validate the prediction performance of the log-Gabor method in four ways: (1) with and without gender information, (2) training on the training set and testing on the development set, (3) by 10-fold cross-validation on the training and development sets and (4) training on both the test and development set and testing on the test set. For this last experiment we used the parameters of the best performing model in the cross-validation experiment and retrained the SVM with all labeled instances for prediction of the labels of the Test instances.

4. Log-Gabor results

In this section we first present the results on the labeled parts of the datasets, i.e. training on the Training sets and testing on the Development sets and 10-fold cross validation on both the Training and Development sets, and then present the results for the Test sets.

The performance parameter for evaluation is, as prescribed by the Challenge guidelines, the unweighted average recall, i.e., the average of percentage of correctly classified positive instances and the percentage of correctly classified negative instances.

Table 2. *Results of the method using the INTERSPEECH 2012 Challenge Personality and Likability datasets. The numbers indicate the unweighted average recall of classification. Dev indicates testing on the development set, CV indicates 10-fold cross-validation on the combined training and development set.*

Task	Baseline (%)	Best no gender (%)		Best gender (%)	
	Dev	Dev	CV	Dev	CV
Likability	**58.5**	**67.7**	**64.5**	**74.2**	**67.5**
Openness	60.4	67.8	66.1	73.2	70.4
Conscientiousness	74.9	76.6	76.9	80.3	78.7
Extraversion	82.8	85.3	85.4	89.1	85.6
Agreeableness	67.6	72.0	74.3	72.7	76.3
Neuroticism	68.9	74.2	74.4	75.0	77.5
Average	**70.9**	**75.2**	**75.4**	**78.1**	**77.7**

4.1 Results for labeled sets

In Table 2 we reproduce the baselines of the Sub-Challenges of INTERSPEECH 2012 (second column), which we take to be the maximum of the reported baselines reported in [8], and present the best results obtained with the log-Gabor method per classification (i.e., the binary classification of Likability and the five binary classifications of the Personality dimensions). The table presents the performances for the original dataset with male and female speakers (the column labeled "Best no gender") and for the dataset manually separated into female and male speakers ("Best gender"). The best gender performance is calculated as the weighted mean of the best performances on the female and male subsets. The columns labeled "Dev" and "CV" list the unweighted average recall percentages for evaluation on the development set and using 10-fold cross-validation, respectively.

Overall, these results show that the log-Gabor method performs above baseline, with gender separation having a positive effect on performance. The gender separation results were obtained using perfect knowledge about the gender of the speakers. In a separate gender-classification experiment, we determined how well the log-Gabor method could estimate the gender from the speech samples. It turned out that the method has a gender-estimation accuracy of 96.1% and 99.6% for the Likability and Personality datasets, respectively (using 10-fold cross-validation).

Table 3. *Δt and PCA parameters for the best performance on the development set. An * indicates logarithmic representation of the energy in the filtered images.*

Classification	No gender		Male		Female	
	Δt	PCA	Δt	PCA	Δt	PCA
Likability	5	100	20	60	5.5*	11
Openness	5*	12	3.5	20	11	8
Conscientiousness	30*	10	4.5*	6	5.5*	30
Extraversion	5	100	15	30	2.5	12
Agreeableness	0.5	-	0.5	-	5.5	11
Neuroticism	2.5	12	2.5	11	11	20

Table 3 lists the parameter values associated with the best performances in the cross-validation experiments reported in Table 2. Evidently, the parameter values depend on the classification task. This raises the question how sensitive the data is to change of the parameter values. We have examined the parameter sensitivity for all classification tasks and found that it can vary considerably.

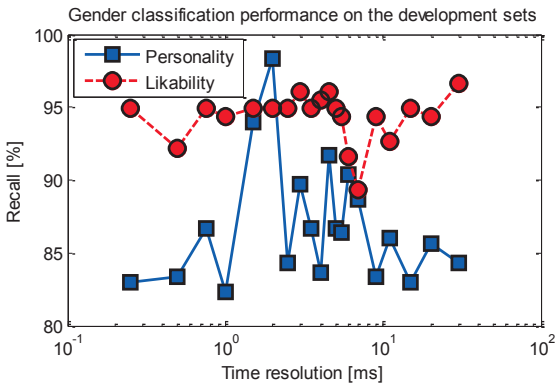

Figure 2: *Illustration of parameter sensitivity using the performances on the development sets of Personality and Likability with varying time resolution (Δt).*

As an illustration, Figure 2 shows the parameter sensitivity for the aforementioned gender classification experiment (tested on the development set). The figure shows that in this experiment the Likability dataset gender classification is less sensitive to variations in Δt than in the Personality dataset. Another illustration is given in Figure 3 which shows the classification performance on the Likability classification task as a function of the number of principal components retained.

Increasing the number of components beyond a minimal number (in this case 3) results in no major improvement of the performance.

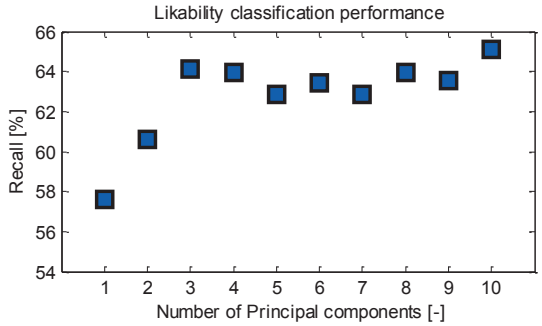

Figure 3: *Performance of the log-Gabor method in the 10-fold cross-validation experiment as a function of the number of principal components.*

4.2 Results for Test sets

Table 4 shows the baselines and results for the Test sets. In the column Best we present the maximum performance over the approaches with and without gender separation.

The results show a considerable drop in performance for all traits. We suspect that our grid search (parameter optimization) gave rise to overfitting. The results for Likability and Neuroticism are above baseline for the Test set and below baseline for the other traits.

Table 4. *Results of the method for the INTERSPEECH 2012 Challenge Personality and Likability Test sets.*

Task	Baseline (%)	No gender (%)	Gender (%)	Best (%)
Likability	**59.0**	**57.8**	**61.4**	**61.4**
Openness	59.0	54.5	48.0	54.5
Conscientiousness	80.1	68.1	76.0	76.0
Extraversion	76.2	67.1	73.4	73.4
Agreeableness	64.2	62.1	62.0	62.1
Neuroticism	65.9	68.2	66.0	68.2
Average	**69.1**	**64.0**	**65.1**	**66.8**

5. Discussion and conclusion

Using the log-Gabor method for speech analysis, we obtained above-baseline performances on the INTERSPEECH 2012 Likability and Personality Sub-Challenges for the labeled sets. Comparable results were achieved for another Likability and Personality dataset (see [12] for more details). Although the drop in performance for the Test sets is rather disappointing, the probable cause (grid search parameter optimization) is easily dealt with. We performed a grid search through parameter space and selected the parameters associated with the best cross-validation performances obtained. From a statistical perspective, this results in an overestimate of the generalization performance, because we use the evaluation performances for optimizing the classification model. We should have performed the parameter optimization within a cross-validation scheme to avoid overfitting. Despite this shortcoming, we think that the results warrant further investigation into the log-Gabor method as a method for the extraction of perceptual cues from speech.

In our future work, we intend to determine the speech features associated with likability and personality. This requires an inverse mapping from the (successful) predictions of the classification model, to the level of speech. In fact, such identification in terms of features is more useful and informative than a more or less black-box approach. We conclude by stating the log-Gabor method is a promising method for speaker trait classification that deserves further investigation.

6. Acknowledgement

The authors thank Marie Postma for her helpful comments on earlier versions of this paper.

7. References

[1] Ezzat, T., Bouvrie, J., and Poggio, T., "Spectro-temporal analysis of speech using 2-D Gabor filters", Proc. Interspeech, 96:1-4, 2007.

[2] Elliott, T. M., and Theunissen, F. E., "The modulation transfer function for speech intelligibility", PLoS computational biology, 5(3), 2009.

[3] Wu, S., Falk, T. H., and Chan, W.-Y., "Automatic speech emotion recognition using modulation spectral features", Speech Communication, 53(5):768-785, 2011.

[4] Meyer, B. T., and Kollmeier, B., "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition", Speech Communication, *53*(5):753-767, 2011.

[5] Field, D. J., "Relations between the statistics of natural images and the response properties of cortical cells", Journal of the Optical Society of America A, Optics and image science, 4(12):2379-2394, 1987.

[6] Fischer, S., Šroubek, F., Perrinet, L., Redondo, R., and Cristóbal, G., "Self-Invertible 2D Log-Gabor Wavelets", International Journal of Computer Vision, *75*(2):231-246, 2007.

[7] Daugman, J., "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters", Journal of the Optical Society of America A, 2(7): 1160-1169, 1985.

[8] Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., Weiss, B., "The Interspeech 2012 Speaker Trait Challenge", Proc. Interspeech 2012, ISCA, Portland, OR, USA, 2012.

[9] Gabor, D., "Theory of communication. Part 1: The analysis of information", Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering, 93(26):429-441, 1946.

[10] Chang, C. and Lin, C., "LIBSVM : a library for support vector machines", ACM Transactions on Intelligent Systems and Technology, 2(3):27:1-27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[11] Kovesi, P.D., "MATLAB and Octave Functions for Computer Vision and Image Processing", Centre for Exploration Targeting School of Earth and Environment, The University of Western Australia. Software available at: http://www.csse.uwa.edu.au/~pk/research/matlabfns/.

[12] Postma, M., Tsoumani, O., Buisman, H.J., Jansen, S.N., Schouten, N.Y., Postma, E.O., & Lanken, G. van, "Detecting personality from thin slices of speech: Self-assessment versus observer ratings", TiCC TR 2012-001. Available at: http://www.tilburguniversity.edu/research/institutes-and-research-groups/ticc/research-programs/cc/technical-reports/.