DOI: 10.17185/duepublico/41622

Probabilistic Breadth as an Evaluation Measure of Gaussian Mixture Models used for Acoustic Emotion States

Ronald Böck * **Ingo Siegert** * **Andreas Wendemuth** *

* *Cognitive Systems Group, Otto von Guericke University, Universitätsplatz 2, 39106 Magdeburg, Germany (Tel: +49-(0)391-67-50061; e-mail: ronald.boeck@ovgu.de).*

Abstract: The automatic speech recognition and speech-based emotion recognition is based on statistical learning methods which are usually highly tuned. Using the content and emotional information from spoken utterances this provides the opportunity to generate human-machine communication which achieves characteristics of cognitive systems. The systems or recognisers are based on learning methods which are well-known. But, an interpretation or evaluation of such classifiers is usually challenging. Classifiers identify categorical regions in n-dimensional feature spaces by modelling the observation probability by mixtures of multivariate Gaussian densities. For this, we present an approach which allows a more detailed interpretation of the classifier and provides an insight to the method. Our approach is based on the breadth of the resulting Gaussian model which can be generated from the mixture models given by the classifier. We introduce the method and present first results on the EmoDB corpus using a classifier with seven mixtures per emotion. In this exemplary case the classification performance is 64.48% unweighted average recall over all Leave-One-Speaker-Out tests. Investigating the probability models, we draw first conclusions on the characteristics of the Gaussian mixtures applying the breadth as the only parameter.

1. INTRODUCTION

Human-computer interaction recently received increased attention. Besides making the operation of technical systems as simple as possible, one goal is to enable a natural interaction. In this context, research was and is focused on developing easy-to-use interfaces that could be used by experts as well as by novices (Carroll [2013]).

Two researchers heavily related with the human communication theory are F. Schulz von Thun and P. Watzlawick. Schulz von Thun discussed the many aspects of human communication and introduced his "four-sides model". Besides factual information also appeal, relationship, and self-revelation play important roles (cf. Schulz von Thun [1981]). Watzlawick investigated human communication and formulates five axioms (cf. Watzlawick et al. [1967]), where the axiom: "One cannot not communicate" is the most important. Both, Schulz von Thun and Watzlawick et al. emphasised the importance of the non-verbal behaviour. This indicates that Human-Human Interaction (HHI) uses many different channels for communication. Thus, HHI is usually understood as a mixture of speech, facial expressions, gestures, and body postures. These considerations are not only valid for HHI but also for Human-Machine Interaction (HMI), which stimulated research in multimodal interaction in HMI.

Although factual information is in the focus, users also create a relationship with the system (cf. Lange and Frommer [2011]). Thus, it is important to know "how" something has been said in HMI as well. Further motivated by the book "Affective Computing" (cf. Picard [2000]), the vision emerged that future technical systems should provide a more human-like way of interaction while taking into account human affective signals (cf. Picard and Cook [1984]). This area of research has received increased attention since the mid-2000's, as more and more researchers combined psychological findings with computer science (cf. Zeng et al. [2009]). In order to enable technical systems to recognise emotional states automatically, these systems have to measure input signals, extract emotional characteristics, and assign them to appropriate categories. This approach, known from pattern recognition, has been widely used since the 1980s in computer science (cf. Bishop [2011]).

We are aware of the issue that speech recognition and affect recognition from acoustics have been widely investigated over the last decade (cf. e.g. Anusuya and Katti [2009], Anagnostopoulos et al. [2012]). From literature we can see that during the work on speech topics several classifiers which serve as recognisers were established (cf. e.g. Albornoz et al. [2010], Schuller et al. [2011]).

Besides a few other approaches, the most prominent classifiers are Support Vector Machines (SVMs) and Gaussian Mixtures Models (GMMs) (cf. Reynolds et al. [2000], Schuller et al. [2011, 2009]). Though SVMs and GMMs are widely used in the community and in statistical modelling at large, it is hard to get any impression of the current appearance of each model. SVMs are relatively easy to train but the high dimensionality of the final model representation complicates any (human) interpretation. This is mostly due to the high complexity of SVMs by using the kernel trick, that implicitly maps inputs into high-dimensional feature spaces. In contrast, a detailed anal-

DOI: 10.17185/duepublico/41622

ysis of GMMs is difficult due to the high dimensionality of the observed space but the use of Gaussian mixture models allows an understandable interpretation. A GMM is a mixture distribution of weighted Gaussian distributions, to cover different values of the same feature and to represent non-normal distributions (cf. Ray and Lindsay [2005]). Based on the Probabilistic Acoustic Volume (PAV) Profiles, depicting GMMs as contour lines, we used directly the model parameters (mean, variance, weighting), to investigate the resulting models. This direct observation of GMMs, or more specifically the characteristics of the resulting shape of the GMM, could provide an analytical description of the "covered area" A_n. Therefore, the intention is to derive an approach for such an analysis based on the given components of a GMM.

What does "covered area" mean? In this paper, we consider a space which is spanned by various speech features assigned to a particular emotion. It is assumed that each emotional state can be represented in a multi-dimensional space (cf. e.g. Mehrabian [1996], Scherer [2005]). In this space the classifier generates a mapping of the acoustic information onto emotional categories. We do not consider the meaning of features, any selection, or mathematical modelling of proper mapping function. In our approach the classifier and thus, its characteristics are already given. As aforementioned, the classifier selects emotional categories and these are grounded in the multi-dimensional space. In terms of GMMs, they try to fit the observation probability of the corresponding emotional state in the multi-dimensional space, what we will call "covered area". From our point of view, this area provides additional information on the quality of the classification.

Therefore, the goal of this paper is to present an approach to analyse GMMs in terms of the covered area in multi-dimensional space. In general, the method is independent from the task as such and can be, thus, used in various fields like recognition engines and statistical modelling. In this paper, we focus on the acoustic sensor input assuming that an interpretation of this modality will contribute to a more general understanding of a multiple sensor system like a companion (cf. Wendemuth and Biundo [2012]). In particular, we consider the emotional content of spoken utterances. This means, the pure content will be neglected but the way how the utterance is expressed is analysed.

2. METHOD

In this section we introduce an approach to analyse GMMs which are used in automatic classifiers. In this paper the focus is on GMMs generated on emotionally coloured speech. In particular, the analysis is not intended to discuss or debate the utilised features in classification, but provides a method which may help to obtain a better understanding of the resulting classifier. Of course, feature selection is a highly interesting topic in the community. On the other hand, it can be solved only in an interdisciplinary discussion on various levels amongst researchers.

At first, we consider the internal structure of GMM-based classifiers. As shown in Figure 1, such a classifier is usually built from an accumulation of several models. All of them are normal distributed. Hence, the final classification result is based on a weighted mixture of those Gaussian models. In Figure 1 such a mixture model is plotted. In the remaining paper we call the final model the resulting GMM of the classifier. This resulting GMM is in the focus of our analysis and approach.

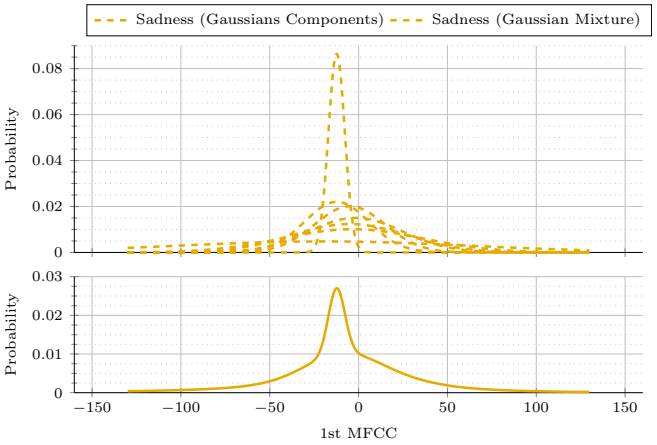

Fig. 1. The plot shows the seven components (top plot) generating the resulting GMM (bottom plot) of the uttered emotion *sadness*, in the first dimension (first MFCC).

Our approach is inspired by the PAV Profile introduced in Cummins et al. [2014], where based on PAV a height profile on several probabilistic steps of the resulting GMM is generated. For this, they achieve an isoclinic-like profile of their applied classifier. In fact, Cummings et al. have to estimate the parameters of their model by a Monte-Carlo approach since the underlying model of the classifier is not given in their work. In Cummins et al. [2014], the PAV Profile is afterwards used as a feature to distinguish and further classify several levels of depressed speech. For their investigation the assumption holds that depressive speech has a broader PAV Profile than non-depressive speech (cf. Cummins et al. [2014]).

In contrast to Cummins et al. [2014], we do neither intend to generate such a profile of the resulting GMM nor to use it as an additional feature for classification, yet. The aim of our approach is to derive a statement on the covered area in the multi-dimensional space of a resulting GMM composed by mixture of various Gaussian models. The covered area (cf. the idea of PAV) is assumed to be correlated with the ability to classify a particular emotional state. Therefore, we look for a measure which indicates the covering of a certain area in a multi-dimensional space and that is also easy to compute. The breadth of a resulting GMM is a good estimator for such covering since it provides an impression of the multi-dimensional space's part which is under the influence of the mixture model.

In the following description of the approach we focus on GMMs with seven components. This is no restriction of the method itself but provides us with the option of a better visualisation of the approach's steps (cf. Figures 1 and 2). In contrast to Cummins et al. [2014], we have full access to the trained models and their parameters during the analysis. Our models are trained with the Hidden Markov Toolkit (HTK) (cf. Young et al. [2009]), a statistical training tool for GMMs by the University Cambridge,

DOI: 10.17185/duepublico/41622

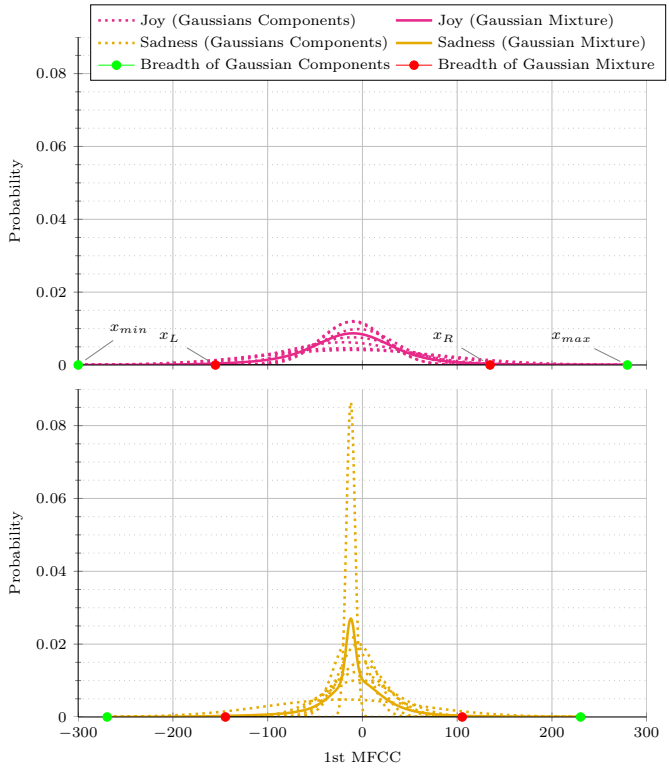

Fig. 2. The plot shows the seven components (dashed lines) generating the resulting GMM (solid lines) of the uttered emotions *joy* (top plot) and *sadness* (bottom plot). The outer markers represent the investigation interval whereas the inner, red markers indicate the arguments fulfilling the $\pm 3\sigma$ criterion. These latter values are used to calculate the GMM's breadth.

which provides the mean, variance, and weighting values of each mixture component. As investigated data set in this paper the EmoDB corpus (cf. Burkhardt et al. [2005] and Section 3.1) was used.

At first, the term "breadth" should be clarified. We define the breadth of an GMM as the length of the range where most of the measurements are covered by the corresponding probability distribution. In fact, we will neglect outliers which are too far away from the mean value of a given model. From stochastics we know that a good interval can be found looking at multiples of the variance. Given a Gaussian distribution, an interval of $\pm 3\sigma$ from mean covers approximately 99.73% of all measurements along one feature dimension (cf. Section A). Therefore, we define according to Equation (1) the breadth b of a resulting GMM as the distance between the arguments of $x_l = \mu - 3\sigma$ and $x_r = \mu + 3\sigma$ calculated on the distribution of the resulting GMM (cf. inner markers in Figure 2).

$$b = x_r - x_l \tag{1}$$

Given this definition, we still have to find the corresponding arguments x_l and x_r. For searching we can define an investigation interval which is fixed by the components of the GMM. In fact, we construct the interval utilising the smallest -3σ value of all components as lower bound $x_{\min}$ and the largest $+3\sigma$ value as upper bound $x_{\max}$. The assumption that the breadth of the resulting GMM is in $[x_{\min}, x_{\max}]$ holds true since the respective GMM is generated by a weighted sum of all components with weights ≤ 1.0 each (cf. Figures 1 and 2). Further, the resulting GMM is in between the generating mixtures with the largest and lowest weight (identity is not excluded). Therefore, it is to notice that in general $x_{min} \leq x_l$ and $x_r \leq x_{max}$.

For our analyses, we investigate each feature separately looking for characteristics in the breadth of the corresponding GMM. In the setup of GMMs, very typically diagonal covariance matrices are being used (cf. e.g. Young et al. [2009]), in order to keep the number of variable parameters linear with the (high) dimensionality of the problem. In the case of general (non-diagonal) covariance matrices, projections on a subspace (e.g. 1-dimension) can be achieved by explicit computation of marginalisations which is easily done for multivariate normal distributions $\mathcal{N}(x|\mu, \Sigma)$ (cf. e.g. Do [2008]). Therefore, we are looking for a marginalisation of the general probability distribution. In the case of a non-diagonal covariance matrix the marginalisation for a dimension m is done as follows:

$$G(x^m) = \sum_i w_i \, \mathcal{N}(x|\mu_i^m, \Sigma_i^{(m,m)}), \tag{2}$$

where w_i is the weight of the i^{th} component. For each component yields

$$\int_{-\infty}^{+\infty} \mathcal{N}(x|\mu, \Sigma) dx = 1 \tag{3}$$

and this yields

$$\sum_i w_i = 1. \tag{4}$$

Using this kind of marginalisation, it is to notice that the original distribution cannot be regenerated by the product of the marginalised probability distributions.

In general, given the marginalised probability, we are searching for the corresponding arguments of the resulting GMM that fit the $\pm 3\sigma$ criterion. In general, this would be done by analysing and solving the integral equation with the intended result of 0.9973. Unfortunately, the integral of the Gaussian distribution cannot be computed directly, but the corresponding error function (cf. Equation (5)) is helping us.

$$\begin{aligned} \operatorname{erf}(y) &= \int_{-\infty}^{y} \mathcal{N}(x) dx \\ &= \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \end{aligned} \tag{5}$$

To obtain the arguments fitting the 3σ criterion we will apply an iterative search.

At first, we calculate (in one dimension) the argument x_r which is the upper bound on the GMM's breadth. For this,

$$\begin{aligned} \int_{-\infty}^{x_r} p(x) dx &= \sum_i w_i \int_{-\infty}^{x_r} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\frac{(x-\mu_i)^2}{\sigma_i^2}} dx, \\ &\quad \text{replacing by } z = \frac{x - \mu_i}{\sigma_i} \\ &= \sum_i w_i \int_{-\infty}^{\frac{x_r - \mu_i}{\sigma_i}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \\ &= \sum_i w_i \operatorname{erf}(\frac{x_r - \mu_i}{\sigma_i}), \end{aligned} \tag{6}$$

DOI: 10.17185/duepublico/41622

where μ_i and σ_i are the mean and variance of the i^{th} component, respectively. Further, we know

$$\begin{aligned}\int_{-\infty}^{3\sigma} \mathcal{N}(x|0,\sigma)dx &= \int_{-\infty}^{3} \mathcal{N}(x|0,1)dx \\ &= \text{erf}(3).\end{aligned} \tag{7}$$

Since we are looking for the corresponding argument, we combine Equations (6) and (7) and obtain

$$\begin{aligned}G(x_r) &= \sum_i w_i \ \text{erf}(\frac{x_r - \mu_i}{\sigma_i}) \\ &\stackrel{!}{=} \text{erf}(3) \\ &= 0.99865.\end{aligned} \tag{8}$$

From Equation (8), x_r will be found by an iterative 1-dimensional search in $[x_{\min}, x_{\max}]$.

Similar to the calculation of x_r, the lower bound x_l is derived. The important difference is that the left bound $1 - \text{erf}(3)$ is applied for operation.

$$G(x_l) = 1 - \text{erf}(3). \tag{9}$$

Hence, we end up with two arguments x_r (cf. Equation (8)) and x_l (cf. Equation (9)) and thus, can compute the breadth of the resulting GMM as given in Equation (1).

The covered N-dimensional area A_N will be obtained as a product over dimensions (cf. Equation (10)).

$$A_N = \prod_{n=1}^{N} (x_{r_n} - x_{l_n}) \tag{10}$$

In Figure 2 the components, the resulting GMMs, and the markers for the investigation interval as well as the breadth are shown for the uttered emotions *anger* and *neutral*.

3. EXPERIMENTAL RESULTS

In this section, the utilised data set and the classifiers are briefly introduced. Further, results of our approach are discussed.

3.1 EmoDB Corpus

The EmoDB corpus (cf. Burkhardt et al. [2005]), also called Berlin Database of Emotional Speech, is hosted by the Technical University of Berlin. The data set contains 493 high quality audio samples (cf. Schuller et al. [2009]) - each of them last for $1 - 2$ seconds - with German spoken utterances by native speakers. All recordings are done in an anechoic chamber. In total ten actors (five female, five male) performed ten sentences with emotional colouring. These emotions are *anger*, *boredom*, *disgust*, *fear*, *joy*, *neutral*, and *sadness*. The sentences' content is not related to the expressed emotions. To ensure a high quality, various listeners' tests were applied to the material by Burkhardt et al. (cf. Burkhardt et al. [2005]).

In our experiments, we applied a speaker-independent evaluation scheme. For this, we need speech samples of each emotion from all speakers. In EmoDB for *disgust* material is not available for all speakers. Therefore, *disgust* is neglected in our investigation. Hence, the classifiers select one emotional class out of six, where six classifiers were trained in a "1-against-all" mode.

3.2 Trained Classifiers

For this paper, we investigated GMMs with only seven mixture components. This is no restriction of the method but increases the readability of the visualisation. From previous experiments we know that depending on the emotions different number of mixtures can be used resulting in almost similar recognition performance (cf. Schuller et al. [2009], Böck et al. [2013], Siegert et al. [2014], Vlasenko et al. [2014]).

The applied features are selected corresponding to Böck et al. [2010], namely Mel-Frequency Cepstral Coefficients (MFCCs) 1 to 12 and the zeroth cepstral coefficient with the respective delta and acceleration values, resulting in a 39-dimensional feature space. The whole classifier was trained on EmoDB (cf. Section 3.1) using HTK (cf. Young et al. [2009]). For each emotion a separate GMM with individual mixture components was trained. For this, we used a Leave-One-Speaker-Out (LOSO) strategy. In LOSO, the samples of one speaker were only used for testing and all remaining material of the other speakers is applied for training, leading to a classifier which has to generalise to achieve a high performance since no material of the test person is seen during training. All utterances are windowed with a Hamming window of window size 25ms and an overlap of 10ms.

Tab. 1. Unweighted Averaged Recall (UAR) for using each speaker as test for a LOSO validation on EmoDB.

Speaker	UAR [%]
03	89.88
08	80.83
09	54.27
10	54.88
11	66.23
12	53.89
13	81.11
14	71.08
15	71.61
16	65.85

The individual unweighted average recall for each speaker is gevien in Table 1. The averaged unweighted average recall of the classifier on LOSO tests is 68.96% with a standard deviation of 12.45%. For the discussion and plotting we focus on the emotion classifier which achieved the best results (Speaker 03), assuming that this shows the best generalisation. The unweighted average recall of the particular classifier is 89.88%.

3.3 Results

Discussing the results, we focus on the resulting GMM for each emotion. How this is constructed is described in Section 2 and visualised in Figures 1 and 2. An overview for breadth of the 1st MFCC for each individual speaker is depicted in Figure 3. In Figures 4 to 6 the corresponding resulting GMMs are shown for the first to third MFCCs.

From the breadth values and the plots we can see that we can distinguish three major groups of GMMs as "narrow", "medium", and "wide" breadth. For instance in Table 2

DOI: 10.17185/duepublico/41622

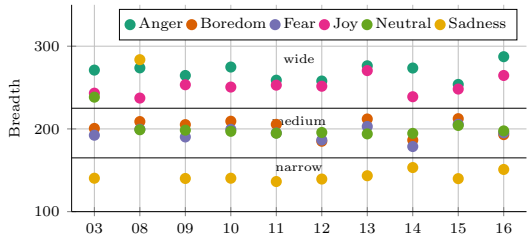

Fig. 3. The breadth values for the 1st MFCC of all trained GMMs of all speakers on emoDB.

the corresponding breadth values of MFCC 1 for each emotion are listed.

Tab. 2. Mean breadth value over al speakers grouping of MFCC 1 feature for each emotion in EmoDB.

Group	Emotion	Mean Breadth Value
narrow	sadness	156.8
medium	fear	194.6
	neutral	201.4
	boredom	201.9
wide	anger	269.2
	joy	251.1

Unfortunately, for most of the emotions the mentioned grouping (cf. Table 2) keeps not constant comparing the features, except for *joy* that has a wide breadth value. At the moment, we did not define a common threshold, over all investigated features, for the identified groups since more detailed analyses are necessary to derive reasonable threshold values which is definitely in the focus of further research. Using fixed thresholds might lead to a more stable clustering.

On the other hand, analysing the different features in comparison, the principle shaping form is similar (cf. e.g. *sadness* in the various plots). In fact, with EmoDB we investigate a corpus which presents highly expressive emotions with high quality but this allows us to concentrate on the characteristics of the method. Further, Figure 2 indicates that the breadth value reflects it properly. Additionally, we can conclude that a narrow breadth results from a packed combination of generating mixtures.

Tab. 3. Grouping of emotions based on the breadth value for MFCC 1 (cf. Table 2) presenting the peak probability value for the resulting GMM.

Group	Emotion	Mean Peak Value
narrow	sadness	0.0270
medium	fear	0.0150
	neutral	0.0171
	boredom	0.0123
wide	anger	0.0114
	joy	0.0087

Furthermore, the investigation shows that GMMs with smaller breadth ("narrow" group) tend to have higher peak probability values than "wide" breadth GMM's which holds true over all features (cf. Figures 4 to 6). The peak values for MFCC 1 are given in Table 3. In

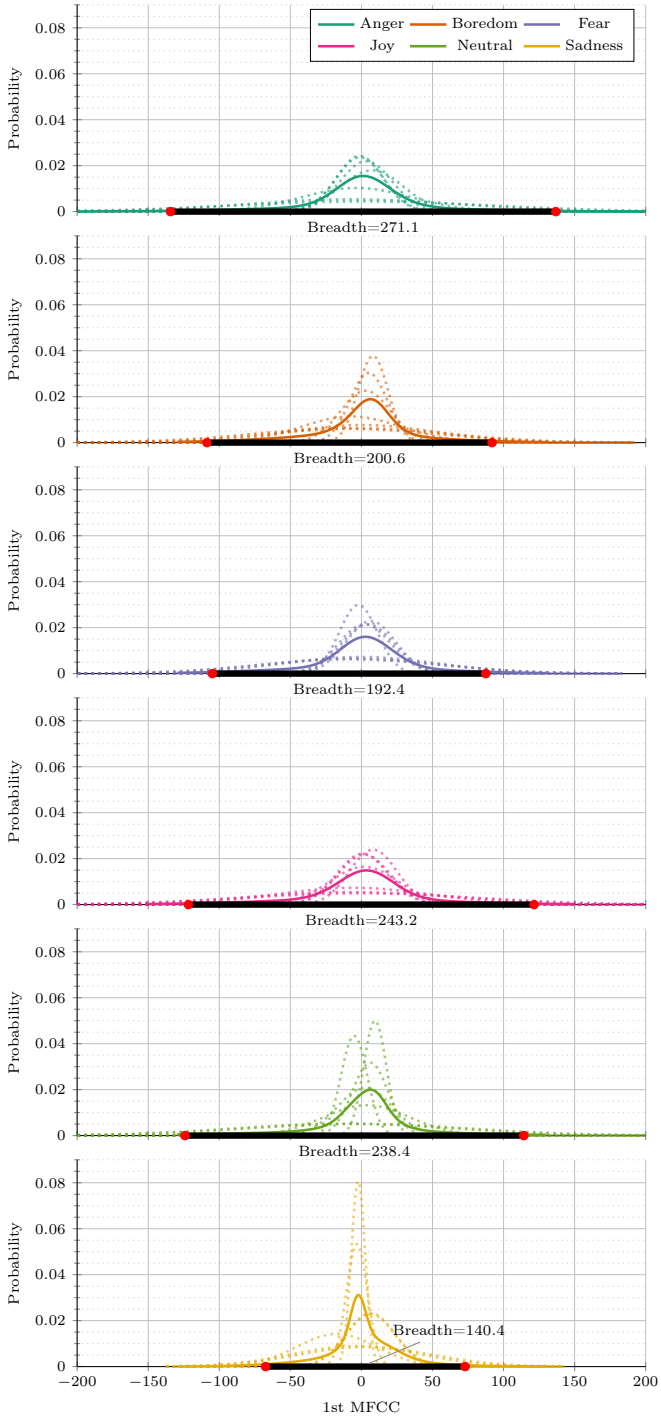

Fig. 4. Comparison of the first MFCC for each emotion with respective breadth. The mixture components as well as the resulting GMMs are plotted in each case. The red markers indicate the arguments fulfilling the $\pm 3\sigma$ criterion. Be aware of the different scaling for *sadness*.

other words, broader resulting GMMs have a trend towards shallow curves which means, that both features are indirectly proportional. From this relation we can derive statements that can be used to interpret and evaluate a trained classifier. Further, we can see that generating mixtures with high peaks and high weights are fit in the resulting GMM (cf. Figure 2). In this case, *sadness* is a

DOI: 10.17185/duepublico/41622

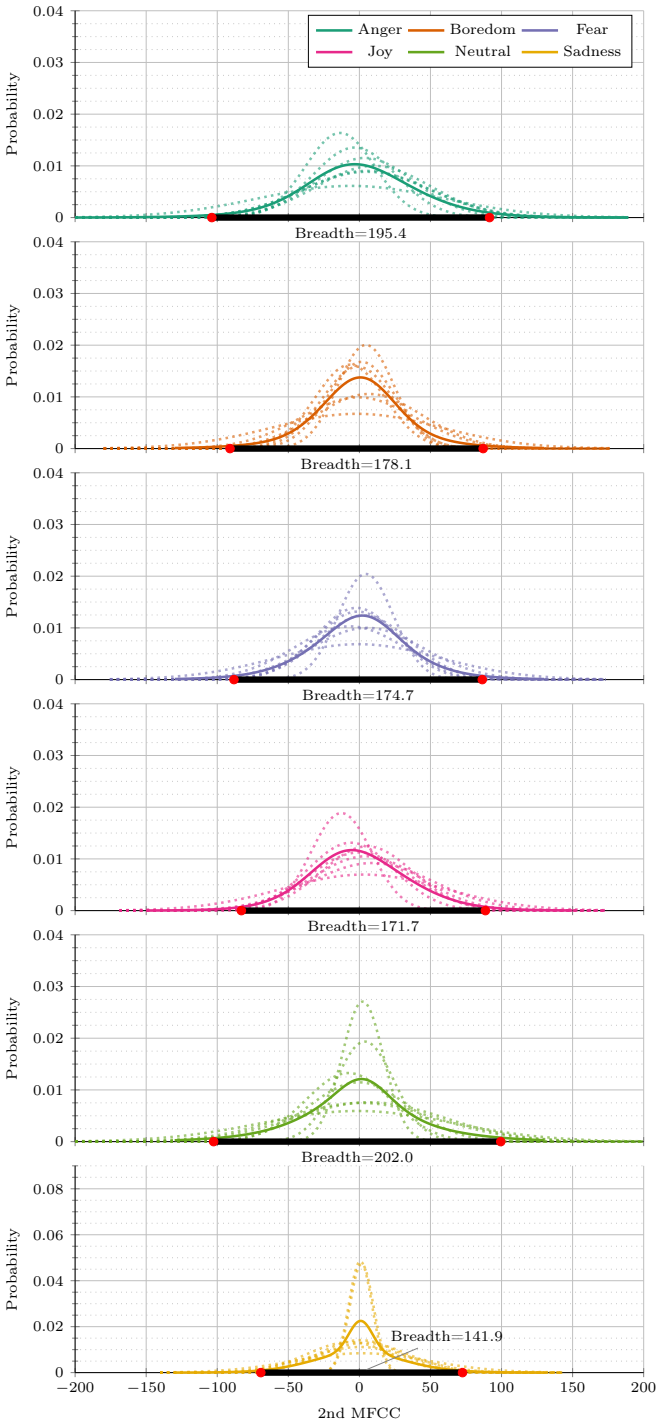

Fig. 5. Comparison of the second MFCC for each emotion with respective breadth. The mixture components as well as the resulting GMMs are plotted in each case. The red markers indicate the arguments fulfilling the $\pm 3\sigma$ criterion. Be aware of the different scaling for *sadness*.

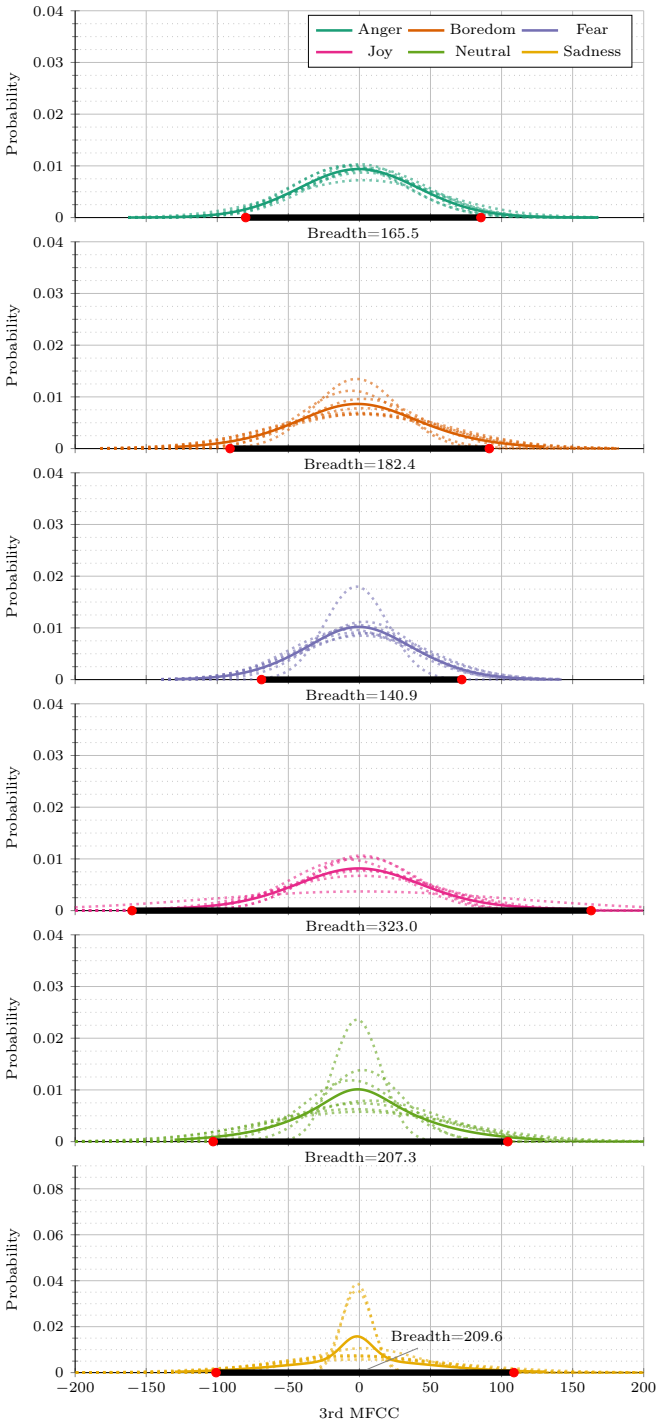

Fig. 6. Comparison of the third MFCC for each emotion with respective breadth. The mixture components as well as the resulting GMMs are plotted in each case. The red markers indicate the arguments fulfilling the $\pm 3\sigma$ criterion. Be aware of the different scaling for *sadness*.

good example for the 1st MFCC and the 2nd MFCC. But for the remaining feature, the 3rd MFCC, the grouping changes (cf. Figures 6). Since the breadth is calculated on the $\pm 3\sigma$ criterion we do not have clustered breadth values as in Cummins et al. [2014]. This can be the case if more highly weighted peaks occur. Nevertheless, the resulting GMMs would be influenced by these mixtures given the constructive method discussed in Section 2.

4. CONCLUSION

Inspired by the idea of PAV Profile (cf. Cummins et al. [2014]), we proposed the probabilistic breadth of GMMs

DOI: 10.17185/duepublico/41622

to derive statements on their covered area. This area in a multi-dimensional space is under the influence of the corresponding GMM. The resulting GMM and thus, its covered area is generated by a combination of weighted mixture components. As explained in Section 2, we utilised the arguments of the resulting GMM, fulfilling the $\pm 3\sigma$ criterion, to calculate the breadth of it. For this, the majority of measurements or samples is covered and we can thus provide more detailed insights on the discriminative power of the obtained mixture models. In this paper, we focussed on a few generating mixtures. Nevertheless, the method as such is flexible and could handle any number of mixture components.

In general, the approach is task independent and hence, can be used to evaluate GMMs in any classification domain. We do not discuss any idea of feature selection or feature mapping from an input space to a feature space, but with the presented method, an interpretation of features in terms of a final classifier can be provided. It can be used to visually inspect the obtained components of a classifier and further, help to understand their power. Finally, the approach interprets the differences of resulting GMMs by providing a single number reflecting the narrowness of each GMM in a certain dimension of a multi-dimensional space.

5. FUTURE WORK

Using the method of a GMM's probabilistic breadth for evaluation is a quite universal approach. For this, it can be used in any domain and any case of GMM-based classifiers. For pure analyses of speech, this method may provide insights on the balancing of phonemes or similar effects. On the other hand, the investigation of emotional speech is of more interest since the acoustical insights are still under investigation (feature are not finally fixed, etc.). Therefore, in further research, we will use the presented method to analyse, especially, affective speech recorded in a naturalistic interaction. A later goal is to use the breadth as an additional measure for distinguishing the various affective classes during clustering.

As we already briefly discussed in Section 3.3, the height is related to the breadth of a resulting GMM with packed densities. Further, the shape of the GMM can provide an additional information to the characteristics of covered area. The shaping information can be obtained by a height profile of the GMM (cf. Cummins et al. [2014]; though, the authors did not consider the breadth). Such a height profile is a good combination with the breadth value. Therefore, in future work, the combination of both approaches will lead to more detailed understanding of the utilised classifiers. Additionally, the method will help to achieve more insights in the characteristics of the applied features, also for emotion recognition from speech.

ACKNOWLEDGEMENTS

We acknowledge continued support by the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" (www.sfb-trr-62.de) funded by the German Research Foundation (DFG).

REFERENCES

E.M. Albornoz, D.H. Milone, and H.L. Rufiner. Multiple feature extraction and hierarchical classifiers for emotions recognition. In Anna Esposito, Nick Campbell, Carl Vogel, Amir Hussain, and Anton Nijholt, editors, *Development of Multimodal Interfaces: Active Listening and Synchrony*, volume 5967 LNCS of *Lecture Notes in Computer Science*, pages 242–254. Springer, 2010.

C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, pages 1–23, 2012. without issue assignment.

M. A. Anusuya and S. K. Katti. Speech recognition by machine: A review. *Int. Journal of Computer Science and Information Security*, 6:181–205, 2009.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, Berlin, Heidelberg, Germany, 2 edition, 2011.

R. Böck, D. Hübner, and A. Wendemuth. Determining optimal signal features and parameters for hmm-based emotion classification. In *Proceedings of the 15th IEEE Mediterranean Electrotechnical Conference*, pages 1586–1590, Valletta, Malta, 2010. IEEE.

R. Böck, K. Limbrecht-Ecklundt, I. Siegert, S. Walter, and A. Wendemuth. Audio-based pre-classification for semi-automatic facial expression coding. In M. Kurosu, editor, *Human-Computer Interaction*, volume 8008 of *Lecture Notes in Computer Science*, pages 301–309. Springer, 2013.

F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A database of german emotional speech. In *Proc. of the INTERSPEECH-2005*, pages 1517–1520, Lisbon, Portugal, 2005.

J. M. Carroll. *Human Computer Interaction - brief intro*, page s.p. The Interaction Design Foundation, Aarhus, Denmark, 2 edition, 2013.

N. Cummins, V. Sethu, J. Epps, and J. Krajewski. Probabilistic acoustic volume analysis for speech affected by depression. In *INTERSPEECH-2014*, pages 1238–1242, Singapore, 2014. ISCA.

C.B. Do. More on multivariate gaussians. Technical report, Stanford University, 2008.

J. Lange and J. Frommer. Subjektives Erleben und intentionale Einstellung in Interviews zur Nutzer-Companion-Interaktion. In *Proceedings der 41. GI-Jahrestagung*, volume 192 of *Lecture Notes in Computer Science*, pages 240–254. Bonner Köllen Verlag, Berlin, Germany, 2011.

A. Mehrabian. Analysis of the big-five personality factors in terms of the pad temperament model. *Aust J Psychol*, 48:86–92, 1996.

R. R. Picard and R. D. Cook. Cross-validation of regression models. *J Am Stat Assoc*, 79:575–583, 09 1984.

R.W. Picard. *Affective computing*. MIT Press, Cambridge, USA, 2000.

R. Ray and B. Lindsay. The topography of multivariate normal mixtures. *he Annals of Statistics*, 33:2042–2065, 2005.

D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digit Signal Process*, 10:19–41, 2000.

K. R. Scherer. *Unconscious Processes in Emotion: The Bulk of the Iceberg*, pages 312–334. Guilford Press, New

DOI: 10.17185/duepublico/41622

York, USA, 2005.

B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth. Acoustic Emotion Recognition: A Benchmark Comparison of Performances. In *Proc. of the IEEE ASRU-2009*, pages 552–557, Merano, Italy, 2009.

B. Schuller, A. Batliner, S. Steidl, and D. Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Commun*, 53:1062–1087, 11 2011.

F. Schulz von Thun. *Miteinander reden 1 - Störungen und Klrungen*. Rowohlt, Reinbek, Germany, 1981.

I. Siegert, D. Philippou-Hübner, K. Hartmann, R. Böck, and A. Wendemuth. Investigations on speaker group dependent modelling for affect recognition from speech. *Cognitive Computation*, 6(4):892–913, 2014.

B. Vlasenko, D. Prylipko, R. Böck, and A. Wendemuth. Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications. *Computer Speech and Language*, 28(2):483–500, 2014.

P. Watzlawick, J. H. Beavin, and D. D. Jackson. *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies, and Paradoxes*. Norton, Bern, Switzerland, 1967.

A. Wendemuth and S. Biundo. A companion technology for cognitive technical systems. In A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, and V. C. Müller, editors, *Cognitive Behavioural Systems. COST 2102.*, volume 7403 LNCS, pages 89–103, Dresden, Germany, 2012. Springer.

S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book, version 3.4*. Cambridge University Engineering Department, 2009.

Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31:39–58, 2009.

Appendix A. COVERAGE OF $\pm 3\sigma$

We assume that the random variable X is normal distributed

$$X \sim \mathcal{N}(\mu, \sigma). \tag{A.1}$$

For this, we obtain

$$\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1). \tag{A.2}$$

The goal is to show that $P(X \in [\mu-3\sigma, \mu+3\sigma]) = 0.9973$. Therefore, we can calculate

$$\begin{aligned} P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &= P(-3 \leq \frac{X-\mu}{\sigma} \leq +3) \\ &= \operatorname{erf}(3) - \operatorname{erf}(-3) \\ &= 2\ \operatorname{erf}(3) - 1 \\ &= 0.9973, \end{aligned} \tag{A.3}$$

where $\operatorname{erf}(3) = 0.99865$ as given in corresponding tables.□