APPLYING THE SPEAKING RATE IN A HIERARCHICAL CLASSIFIER FOR EMOTION RECOGNITION FROM SPEECH

David Philippou-Hübner, Ronald Böck and Andreas Wendemuth

Otto von Guericke University Magdeburg
Department of Electrical Engineering and Information Technology
Chair of Cognitive Systems
david.huebner@ovgu.de

Abstract: Humans can easily estimate the rate of speech of a dialog partner in a conversation. Hence, the speaking rate can be regarded as a quite obvious prosodic characteristic of human speech. In particular, it provides information on different emotional dispositions of our dialog partners. However, most machines still lack of such human abilities and therefore research activities have started to focus stronger also on the emotional aspect of speech. In this paper we introduce a hierarchical classifier for emotions from speech. In a two step approach first a binary classification in low and high arousal emotions takes place on basis of the speaking rate feature. Afterwards, a second classification step determines the actual emotion. The hierarchical classifier consists of three Multi-Layer Perceptrons (MLP) trained on cepstral turn-level features, while the speaking rates are determined by applying a broad phonetic class recognizer. We present the results on the emotionally expressive EMO-DB corpus and compare them with results from a single MLP representing a flat approach with no hierarchical structure. An increase of accuracy up to 3.0% in certain emotion categories is reported.

1 Introduction

1.1 Determination of the speaking rate

Exploiting information about the emotional state of a user, machines can be enabled to adapt their dialog strategy, depending on the emotion of the user and hence react in a more appropriate and empathic manner. While ongoing research of other groups often bases on pooling together huge numbers of (high level) features in order to fully exploit the feature space our approach however aims stronger at providing a rather small feature set with competitive performance. Especially small devices with low computational power (smart phones) benefit from such a sparse approach.

In order to take advantage from the correlation between a person's emotional state and his speaking rate, one has to find an automated robust method for speaking rate determination. Humans provide this ability at least on a subjective level using categories like slow, normal or fast. An automated recognition of the speaking rate however is a challenging task. Typically estimated from samples of connected speech uttered spontaneously or read out, it is supposed to reflect the speed at which a person executes articulatory movements for speech production ([4], [5]). As a unit of measurement often either words per minute (wpm) or syllables per minute (spm) are used. Especially, the *wpm* measure suffers from varying numbers of syllables in the word [3]. This is problematic in languages like German where compounding of substantives theoretically can generate arbitrary long words. Hence nowadays, *spm* is widely used for measuring

speaking rates. Given a speech sample it is defined as the number of syllables divided by the duration of the sample. Even if this measure can be regarded as more language independent than *wpm* it provides problems with respect to the intrinsic duration of syllables.
Engineering solutions approach the problem mainly from two directions. Exploiting energy and periodic measures by applying convex-hull algorithms syllables can be detected directly on the speech signal [14]. Including the first spectral moment of full-band energy and compressed sub-band energy correlation a robust syllable detection algorithm was introduced by Morgan and Fosler-Lussier [8]. Further, Heinrich [6] presents a method on rhythmicity features, where the speaking rate is correlated to peaks in the short-time energy envelope of the speech signal.
The second approach follows automated speech recognition (ASR) procedures and utilizes the duration of recognized phonemes for speaking rate estimation. Especially, broad phonetic class recognizers which can distinguish between groups of vowels and consonants have been proven to be suitable for speaking rate estimation [10]. Due to the small number of broad phonetic classes (typically 6-8), they are less sensitive with respect to recognition errors than single phoneme recognizers used in standard ASR systems, which often base on up to 50 classes.
Our approach for speaking rate estimation implements a broad phonetic class recognizer that distinguishes eight phonetic classes. For training we apply the Hidden Markov Toolkit (HTK) [15] on a standard corpus in English language. We define the unit representing the speaking rate as phonemes per second (pps).

1.2 The role of the speaking rate in emotional speech

In the field of multidimensional emotion recognition from speech (for instance in the Valence-Arousal-Dominance space) the speaking rate carries important information about the arousal of a user who is verbally interacting with a machine. Already Murray and Arnott [9] presented emotional voice characteristics for Ekman's basic emotions. Their qualitative results are shown in Table 1. The specifications are based on a comparison of the affective/emotional voice to the neutral voice characteristics. We follow an approach comparable to the one of Koolagudi [7] who presents a two stage emotion recognition approach based on speaking rates. In a first stage active, normal, and passive emotions are separated into three gross classes using Mel Frequency Cepstral Coefficients and prosodic features, while in the second state the single emotions are classified within each class. In contrast to our approach, here the speaking rate does not contribute in form of a single number. Instead it appears in a hidden, rather abstract form represented by a complete feature set.

Table 1 - Speaking rates for Ekman's basic emotions according to [9]

Disgust	**Sadness**	**Joy**	**Neutral**	**Anger**	**Fear**
very much slower	slightly slower	slower or faster	-	slightly faster	much faster

1.3 Structure of the paper

The remainder of the paper is organized as follows: Section 2 introduces the corpora, while results of the broad phonetic class approach are described in Section 3. In Section 4 we apply our speaking rate model on an emotional database and provide emotion specific speaking rates. Section 5 introduces the hierarchical classifier and shows the advantages in comparison to flat classifier. Section 6 summarizes the results and gives a conclusion.

2 Corpora

2.1 RM1 corpus

The DARPA Resource Management Continuous Speech Corpora (RM) [12] consist of digitized and transcribed speech in two main sections, RM1 and RM2. The material is in English language and consists of read sentences modeled after a naval resource management task.

In order to build a universal model we chose the speaker independent material of RM1 which contains 80 speakers, each reading 42 sentences from the RM text corpus. The complete material consists of 2880 files. We selected this database as it contains short utterances lasting in most cases between two and four seconds. Hence, we were able to determine the speaking rates over the complete utterance, as we assume rather constant speaking rates in utterances lasting only a few seconds. Further, we wanted to show that also a smaller corpus is capable of training a high performing system in order to determine speaking rates robustly.

2.2 Berlin Database of Emotional Speech

For affected speech we decided to use the popular studio recorded Berlin Emotional Speech Database (EMO-DB) [2]. This database contains acted emotional speech samples. Ten professional actors (five male and five female) spoke ten German emotionally undefined sentences. In order to provide reliable data in each emotion category 20 evaluators took part in a perception-test. Selecting only those emotions, which provide a level of naturalness not less than 60% and a level of recognizability not less than 80%, we reduced the data to 493 utterances. The single emotions however, are not equally distributed within the material: *anger* dominates with 28% while disgust is underrepresented with only 8%. The other emotions contribute approximately 12% each.

3 Estimation of the speaking rate by applying a Broad Phonetic Class Recognizer

3.1 Method

The core module of the recognizer is represented by a Hidden Markov Model (HMM). Such stochastic models have been established in ASR for a long time [13]. We applied the same features and HMM structure as for a state-of-the-art speech recognizer. Table 2 summarizes the parameters and configuration settings.

Table 2 - Model parameters for feature extraction

Feature type	**Characteristic**
HMM type	3 state left-to-right
Gaussian mixtures	8
Training signal feature	39 MFCC_0_D_A
Sampling rate / Hamming window	10ms / 25ms
Number of filterbanks / cepstral lifters	26 / 22

Implementing a broad phonetic class recognizer, the main difference to a ASR system is the lexicon, which provides transcriptions of all occurring words into their corresponding sequences of phonemes. In our case, however it only contains the mapping of phonemes contained in each broad phonetic class to the class label (compare Table 3). Two additional classes handle silent parts within speech. One detects hardly perceivable short pauses (sp) occurring between single words, while the second class covers longer parts of silence (sil).

Table 3 - Broad Phonetic classes

Phonetic Class	Related Phonemes	Label
Monophthongs	ao aa iy uw eh ih uh ah ax en ae er	mt
Diphthongs	ey ay ow aw oy	dt
Stops + Affricatives	p pd b t td d dd k kd g ch jh	st
Fricatives	f v th dh s z ts sh zh hh	fr
Nasals	m n ng	na
Glides + Liquids	l el r dx y w	li
Short pauses	sp	sp
Silence	sil	sil

Since each phonetic class represents a group of phonemes with similar characteristics, we only lose information about the actual phoneme during the recognition process. However, the information about the presence of a phoneme is still maintained. Hence, within the following considerations we will simply use the term phoneme and do not longer distinguish between phonetic classes and phonemes. The outputs of the HTK tool HVite are further processed in order to count the number of phonemes and to compute the speaking rate on basis of the single durations. Finally, we define the speaking rate as the quotient of the number of phonemes and the total duration. Silent parts are not counted as phonemes, but contribute to the total duration. So, for a given number of phonemes an increase of silence decreases the speaking rate.

3.2 Results of the Broad Phonetic Class Recognizer

To prove the accuracy of our results we compared the reference speaking rates with those of the corresponding model output. Here, we distinguish two types of errors: samples which are recognized as too fast and those which are recognized as too slow with respect to the reference rates. We analyzed these two phenomena separately defining the error as the difference between the measured and the reference rate of speech. Looking at the data used for training 52% (av. error: -1.05 pps, std: 0.61) are recognized as too slow and 38% (av. error: 0.90 pps, std: 0.56) as too fast. Only slightly worse are the results on the test data in which only speakers occur who were excluded from the training material. Here 48% (av. error: -1.16 pps, std: 0.75) of the samples are recognized as too slow and 44% (av. error: 1.03 pps, std: 0.63) as too fast. In all cases the absolute average error is about 1 pps, which corresponds to a relative error of 8.1%. In terms of an average RM1 utterance, which consists of 43 phonemes and lasts about 3.5 seconds, an error of 1 pps results in ± 3.5 extra phonemes with respect to a complete utterance. The error mainly originates from phonemes being either overlooked (deletion error) or added (insertion error) during the recognition process. A further unfavorable influence on the result accounts to an incorrect determination of silent parts. In our model this type of error plays a minor part: in average the total duration of estimated silence varies not more than 5.8% from the true part. For a detailed description of the method and the results compare [11].

4 Analysis of speaking rates in emotional speech

Applying our model for speaking rate estimation on data from the emotionally expressive EMO-DB corpus we found 4 groups which can be distinguished on 95% significance level. Table 4 summarizes average speaking rates for each emotion. Their individual means cover a range between 9.9 pps (sadness) and 16.5 pps (fear) while the intra class standard deviations vary between 1.2 pps (disgust) and 2.5 pps (fear). The results reflect partly those of Murray [9]:

fear shows the highest speaking rates, followed by *joy* and *anger*. *Disgust* and *sadness* show the lowest rates. In contrast to Murray's result *neutral* is slightly faster than *joy* and *anger*. The last column of Table 4 shows the average offsets to the target rates. One sees that the errors are a bit higher than those on the test data of RM1. Especially, active emotions which are uttered faster tend to be classified as too slow. But still, our model is robust with respect to the German language.

Table 4 - Clusters of distinguishable emotions

emotion cluster	**av. speaking rate [pps]**	**av. error [pps] too slow / too fast**
fear	16.5	-1.31 / 1.18
neutral	15.3	-1.13 / 0.95
joy	14.5	-1.25 / 1.06
anger	14.0	-1.20 / 0.95
boredom	13.5	-1.15 / 1.11
disgust	10.5	-1.08 / 1.09
sadness	9.9	-0.97 / 1.12

5 The hierarchical classifier

In this section we introduce a hierarchical MLP classifier for emotion recognition. Beside other classifiers like Support-Vector-Machines or HMMs the use of Artificial Neural Networks is state-of-the-art in emotion recognition [1].

5.1 Applied features

We apply an established feature set, based on Mel Frequency Cepstral Coefficients (MFCCs), including the speaking rate as a further feature. From the EMO-DB material we extracted the energy, the first 12 MFFCs, and the zero-coefficient (F0). Further, for each parameter the corresponding values of the first and second derivative were determined. Hence, the final feature set consists of 42 features on frame-level basis. In order to apply these features in MLP classifiers the following 8 turn-level features on utterance level were computed on each of the 42 frame-level features.

1. mean
2. standard deviation
3. 10th percentile: covering 10% of the observations
4. 25th percentile: covering 25% of the observations
5. 50th percentile: median
6. 75th percentile: covering 75% of the observations
7. 90th percentile: covering 90% of the observations
8. zero-crossings: number of zero-crossings/100 samples

A subset of these features was already successfully applied by Albornoz [1]. We extended this feature set by adding derivatives and concentrating more on percentile values, than on minimum and maximum values. Finally, we end up with a feature set of size 336 plus the speaking rate. As the speaking rate is already a turn-level feature it can be applied directly without further processing.

5.2 General MLP configuraions

All MLP classifiers were implemented with Matlab's Neural Network-Toolbox. Beside the input and output layer all MLPs contain one hidden layer. Applying the full feature set, we found following parameter characteristics to produce optimal results:

- Resilient-Back-Propagation (trainrp) for training
- Fermi function (logsig) for the hidden neurons
- Hyperbolic tangent function (tansig) for the output neurons
- 100 neurons in hidden-layer

The learning rate is automatically adapted. In order to take into account both the influences of varying training/test sets and different weight initializations all presented results are the average of $10,000$ single trials.

5.3 1st Hierarchy: low and high arousal emotions

In this pre-classification step we take advantage from the speaking rate feature. Taking into account the results presented in Table 4 we found out that reducing the 4 groups to 2 yields the best results. The 2 classes can separated more efficiently with respect to the occurring speaking rates. The emotions assigned to each group are shown in Table 5. Within the following considerations we will address the two groups as high arousal *(ha)* and low arousal *(la)*. Although, *anger* is normally rated as *(ha)* slightly better results were achieved with the presented configuration.

Table 5 - Pre-classification

high arousal (ha)	**low arousal (la)**
fear, neutral, joy	anger, boredom, disgust, sadness

Since the speaking rate alone is not capable to provide a satisfactory distinction between the two groups, a sparse subset of the remaining features was determined in order to gain representative classification results. Finally, we applied 12 MFCCs, F0, and the energy in addition. Therefore, the MLP representing the first hierarchy has $14 \times 8 + 1 = 113$ input neurons and 2 output neurons (*la* and *ha*).

5.4 2nd Hierarchy: emotion recognition

After a sample was classified as *la* or *ha* in this step the actual classification of the emotion happens. Therefore, two independent MLPs classify the emotion within each group. For this final classification step we apply the whole feature set, so that both MLPs provide 337 input neurons. The outputs however correspond to the number of emotions in each group, which is 3 for high arousal group and 4 for the low arousal group.

5.5 Results of the MLP

The presented results are maintained from evaluating the accumulated confusion matrix over all trials. Also, the weighted accuracy was computed as the average of the single recognition rates. Within the binary pre-classification step (*ha* or *la*) we obtained an accuracy of 86.7% with a balanced level of confusions.

Table 6 summarizes our final results compared to the results obtained from a baseline MLP providing a flat hierarchy and being trained on the complete feature set. In all categories but the neutral an improvement of the recognition performance can be reported. Especially *anger* and *joy*, which normally tend to be confused easily, profit from the chosen hierarchy structure: since *joy* is in the *ha*-group and *anger* is in the *la*-group they are separated within the pre-classification step and cannot be confused any longer. Neutral emotions however show a huge bandwidth of speaking rates and do not profit from such a hierarchical approach.
Further, our results outperfom those of the hierarchical classifier by Albornoz [1], who reported a classification rate of 66.83% using MLPs on EMO-DB.

Table 6 - Accuracies for the different emotions

emotion	**hierarchical classifier [%]**	**baseline [%]**	**improvement [%]**
fear	**68.95**	68.08	**0.87**
neutral	**76.87**	77.53	**-0.66**
joy	**55.15**	52.11	**3.04**
anger	**92.13**	90.26	**1.87**
boredom	**86.22**	84.84	**1.38**
disgust	**88.34**	87.42	**0.92**
sadness	**91.78**	90.99	**0.79**
weighted accuracy	**79.92**	**78.74**	**1.17**

6 Summary & Conclusion

Analyzing emotion-colored speech we obtained individual speaking rates for different emotions. Exploiting the correlation of speaking rates and shown emotion we presented a hierarchical classifier which improves emotion recognition in average by 1.17%. Further, the achieved average classification rates of 79.92% are competitive with current approaches applying Neural-Network classifiers on acted data (compare [1]).
In future research we want to test our models on corpora, which provide more natural emotions. Also, a normalization of the speaking rate with respect to the age of the speaker is essential in realistic scenarios: younger people tend to talk faster than older ones.

Acknowledgment

This work was supported by the Transregional Collaborative Research Centre SFB/TRR 62 ”Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG). We also acknowledge the DFG for financing our computing cluster used for parts of this work.

References

[1] ALBORNOZ, E. M. ; MILONE, D. H. ; RUFINER, H. L.: Spoken emotion recognition using hierarchical classifiers. In: *Computer Speech and Language* 25(3) (2011), S. 556–570

[2] BURKHARDT, F. ; PAESCHKE, A. ; ROLFES, M. ; SENDLMEIER, W. F. ; WEISS, B. : A Database of German Emotional Speech. In: *Proc. of Interspeech 2005, Lisboa* (2005), S. 1517–1520

[3] CARROL, J. : Problems of measuring speech rate. In: *ERIC Document Reproduction Service No. ED011338* (1967)

[4] CRYSTAL, T. ; HOUSE, A. : Segmental durations in connected-speech signals: Preliminary results. In: *Journal of the Acoustical Society of America* 72 (1982), S. 705–716

[5] CRYSTAL, T. ; HOUSE, A. : Articulation rate and the duration of syllables and stress groups in connected speech. In: *Journal of the Acoustical Society of America* 88 (1990), S. 101–112

[6] HEINRICH, C. ; SCHIEL, F. : Estimating Speaking Rate by Means of Rhythmicity Parameters. In: *Proc. of Interspeech 2011, Florence* (2011)

[7] KOOLAGUDI, S. G. ; KROTHAPALLI, R. S.: Two stage emotion recognition based on speaking rate. In: *International Journal of Speech Technology* 14 (2010), S. 35–48

[8] MORGAN, N. ; FOSLER-LUSSIER, E. : Combining Multiple Estimations of Speaking Rate. In: *Proc. of ICASSP 1998, Seattle* (1998), S. 729–732

[9] MURRAY, I. R. ; ARNOTT, J. L.: Toward the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion. In: *Journal of the Acoustical Society of America* 93(2) (1993), S. 1097–1108

[10] PFAU, T. ; RUSKE, G. : Estimating the speaking rate by vowel detection. In: *Proc. of ICASSP 1998, Seattle* (1998), S. 945–948

[11] PHILIPPOU-HÜBNER, D. ; VLASENKO, B. ; BÖCK, R. ; WENDEMUTH, A. : The Performance of the Speaking Rate Parameter in Emotion Recognition from Speech. In: *Proc. of IEEE International Conference on Multimedia and Expo (ICME) (accepted)* (2012)

[12] PRICE, P. ; FISHER, W. M. ; BERNSTEIN, J. ; PALLETT, D. S.: Resource Management RM1 2.0. In: *Linguistic Data Consortium, Philadelphia* (1993)

[13] RABINER, L. ; BIING-HWANG, J. : Fundamentals of Speech Recognition. In: *Prentice Hall* (1993)

[14] XIE, Z. ; NIYOGI, P. : Robust Acoustic-based Syllable Detection. In: *Proc. of Interspeech 2006, Pittsburgh* (2006)

[15] YOUNG, S. ; EVERMANN, G. ; GALES, M. ; HAIN, T. ; KERSHAW, D. ; LIU, X. ; MOORE, G. ; ODELL, J. ; OLLASON, D. ; POVEY, D. ; VALTCHEV, V. ; WOODLAND, P. : The HTK Book. In: *Cambridge University Engineering Department* (2006)