INTERSPEECH 2009 BRIGHTON

Processing affected speech within human machine interaction

Bogdan Vlasenko, Andreas Wendemuth

Cognitive Systems, IESK, Otto-von-Guericke Universität,
Magdeburg, Germany

bogdan.vlasenko@ovgu.de

10.21437/Interspeech.2009-586

Abstract

Spoken dialog systems (SDS) integrated into human-machine interaction interfaces is becoming a standard technology. Current state-of-the-art SDS, usually, is not able to provide for the user a natural way of communication. Existing automated dialog systems do not dedicate enough attention to problems in the interaction related to affected user behavior. As a result, Automatic Speech Recognition (ASR) engines are not able to recognize affected speech and dialog strategy does not make use of the user's emotional state. This paper addresses some aspects of processing affected speech within natural human-machine interaction. First of all, we propose an affected speech adapted ASR engine. Second, we describe our methods of emotion recognition within speech and present our results of emotion classification within Interspeech 2009 Emotion Challenge. Third, we test affected speech adapted speech recognition models and introduce an approach to achieve emotion adaptive dialog management in human-machine interaction.

Index Terms: Emotion Recognition, Affected Speech Recognition, Emotion Challenge.

1. Introduction

The importance of human behavior based dialog strategies in human machine interaction (HMI) lies in existing limitations of automatic speech recognition technology. Current state-of-the-art Automatic Speech Recognition (ASR) approaches still cannot deal with flexible, unrestricted user's language and emotional prosody colored speech[1]. Therefore, problems caused by misunderstanding a user who refuses to follow a predefined, and usually restricting, set of communicational rules seems to be inevitable.

In the domain of human-machine interaction [2], we witness the rapid increase of research interest in affected user behavior. However, some aspects of affected user behavior during HMI still turns out to be a challenge for developers of Spoken Dialog Systems (SDS).

Detecting and utilizing non-lexical or paralinguistic cues as part of the user behavior state descriptors is one of the major challenges in the development of reliable human-machine interfaces. Notable among these cues are the universal categorical emotional states (anger, happy, sadness, etc.), prevalent in day-to-day scenarios. Knowing such emotional states can help adjust system responses so that the user of such a system can be more engaged and have a more effective interaction with the system [3][4].

The primary aim of this paper is to present our emotion recognition technique and show results of affected speech adapted speech recognition evaluation. Also within this paper, we are presenting results of emotion challenge evaluation, to test our emotion classification engine on huge spontaneous well annotated emotional corpora.

2. Human Machine Interface

Multimodal Human Machine Interfaces (MHMI) has recently become a new feature for different applications [5]. We describe one of possible MHMI architecture for a Spoken Dialog System. Humans employ several output modalities (mimics, speech, prosody) to communicate with a computer. In the Neurobiologically Inspired, Multimodal Intention Recognition for Technical Communication Systems (NIMITEK) modular technology[6], we include recognition of the user's facial expression which is done by a camera and emotion recognition within speech. From speech, obviously commands are recognized. But also, speech and mimics together/separately serve as a multimodal/unimodal emotion source. Emotions are then classified into one of two emotion classes (neutral and anger).

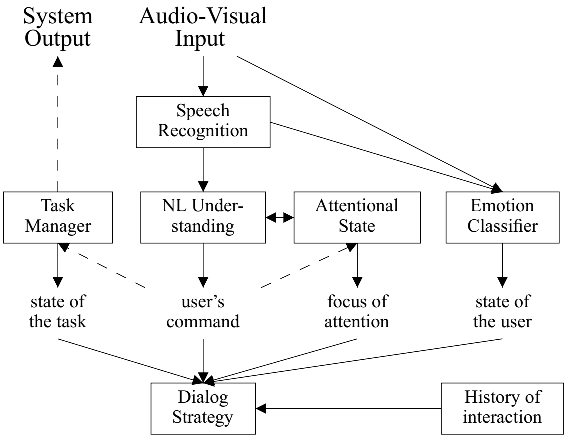

Figure 1: Processing of user's input

Moreover, taking also into account the history of the previous system-user interaction, intentions are classified such as cooperative, explorative or destructive. The latter applies to users who wish to drive the system into dead ends, e.g. by deliberately mispronouncing words or giving contradictory commands. With the recognized commands and the emotional and intentional state, the task controller is driven.

Processing of a user's command in the NIMITEK prototype system is represented in Figure 1. Technical details of Dialog Management Model can be found in [2] and other publications of Gnjatović. Also a more detailed NIMITEK system overview can be found in [7].

3. Affected Speech Processing

3.1. Corpora

To train German monophones we used The Kiel Corpus of Read Speech. The Kiel Corpus is a growing collection of read and spontaneous German which has been collected and labeled segmentally since 1990. The Kiel Corpus comprises over four hours of labeled read speech of 26 female and 27 male speakers. The training set contain 2872 sentences, test set 1001 sentences.

To train an emotional utterance and phoneme models and adapted German monophones ASR models, we decided for the popular studio recorded Berlin Emotional Speech Database (EMO-DB) [8], which covers the 'big six' emotion set (MPEG-4) with boredom instead of surprise, and added neutrality. This database contains acted samples. However, to our best knowledge this is the only public emotional speech database that provides accurate manual syllable boundaries and transcription for model training. 10 (5f) professional actors speak 10 German emotionally undefined sentences (from 6 to 14 words per sentence). 494 phrases are marked as min. 60% natural and min. 80% assignable by 20 subjects. 84.3% accuracy is reported for a human perception test for complete set of emotions and 96.2% for neutral and anger samples subset.

Within the Emotion Challenge we provide some experiments based on FAU AIBO corpus[9]. The whole corpus consisting of 18,216 chunks (9,959 for training and 8,257 for testing) is used for this challenge. For the five-class classification problem, the cover classes Anger (subsuming angry, touchy, and reprimanding) Emphatic, Neutral, Positive (subsuming motherese and joyful), and Rest are to be discriminated. The two-class problem consists of the cover classes NEGative (subsuming angry, touchy, reprimanding, and emphatic) and IDLe (consisting of all nonnegative states).

3.2. Acoustic Features

Speech input for our Dialog Systems is processed using a 25ms Hamming window, with a frame rate of 10ms. As in typical speech recognition we employ a 39 dimensional feature vector per each frame consisting of 12 MFCC and log frame energy plus speed and acceleration coefficients. Cepstral Mean Substraction (CMS) and variance normalization are applied to better cope with channel characteristics.

Within emotion challenge evaluations we tested different feature sets including: 12 MFCC coefficients including delta and accelerations, pitch and energy frame contours. Cepstral Mean Substraction (CMS) and Vocal Tract Length Normalization (VTLN) are tested as an additional option of acoustic feature sets.

3.3. Affected Speech Recognition

As usual for Speech Processing Applications for Spoken Dialog Systems we are using the HTK [10] toolkit for our purpose. For real time Automatic Speech Recognition (ASR) within the Spoken Dialog System, we used ATK. Monophones are modeled by training three emitting state Hidden Markov Model (HMM). A HMM of 3 states and 16 Mixtures of Gaussians was built for each phoneme. We are using the short version of German SAMPA which includes 36 phonemes.

As you can see in Figure 2, affected speech has a different nature in spectral domain. For this reason we decide to test affective speech adaptation technique for robust natural speech recognition. For our Speech recognition evaluation we trained 3 different HMM model sets. First, we trained monophones models on Kiel corpus. Second, on EMO-DB corpus. Third, we applied MAP adaptation for Kiel trained monophones HMM set with EMO-DB samples. This is possible, because in both corpora we have sufficient amount of German monophes samples. For language modeling we apply a domain dependent predefined grammar. The garbage word model encapsulates possible Out of Vocabulary words. In the case of Emotion Challenge, we applied MAP adaptation for Kiel trained monophones HMM adapted by EMO-DB with AIBO samples. Also, word level Bigrams are used for language modeling.

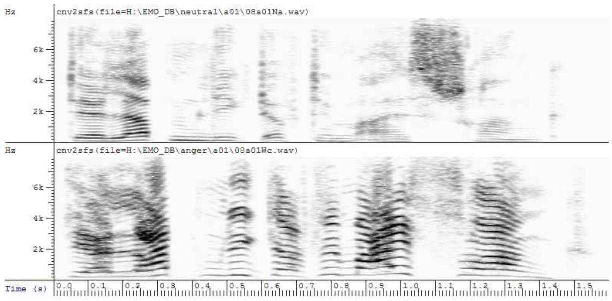

Figure 2: *Spectrogram of anger and neutral speech. Speaker and utterance are identical in both cases.*

3.4. Emotion Classification

For real time emotion classification within speech we used modified HTK and ATK. Two possible classification units of analysis: utterance and phoneme are used[11]. In case of current NIMITEK demonstration system are using real time phoneme level emotion classification.

Within the emotion classification challenge we use a modified HTK toolkit, gender dependent VTLN script and Praat [12] for high-level prosodic features extraction to choose optimal feature sets for two-class and five emotion classification task. Also, whithin emotion challenge evaluation we provide results of combined phoneme and utterance level of analysis.

3.4.1. Utterance Level Analysis

We consider using a speaker recognition system to recognize emotion from speech in the first place. Likewise, instead of the usual task to deduce the most likely speaker (from a known speaker set) Ω_k from a given sequence X of M acoustic observations x, we will recognize the current emotion. This is solved by a stochastic approach following equation 1,

$$\Omega_k = \underset{\Omega}{\operatorname{argmax}} P(\Omega|X) = \underset{\Omega}{\operatorname{argmax}} \frac{P(X|\Omega)P(\Omega)}{P(X)} \quad (1)$$

where $P(X|\Omega)$ is called the emotion acoustic model, $P(\Omega)$ is the prior user behavior information and Ω is one of all emotions known by the system. In case of turn level analysis the emotion acoustic model is designed by s single state HMMs. The states are associated with emission-probabilities $P(X|s)$ which for continuous variables x are replaced with their probability density functions (PDF). These PDFs are realized using weighted sums of elementary Gaussian PDFs (Gaussian Mixtures Models, GMM). Each emotion is modeled by its own GMM. One emotion is assigned for a full dialog turn.

3.4.2. *Phoneme Level Analysis*

As an alternative emotional unit we choose phonemes, as these should provide the most flexible basis for unit-specific models: if emotion and level of interest recognition is feasible on phoneme basis, these units could be most easily re-used for any further content, and high numbers of training instances could be obtained [11].

We use a simple conceptual model of dynamic emotional state recognition on phoneme level analysis: the full list of 36 phonemes is modeled for neutral and anger emotion speaking style, independently. As a result 2 x 36 = 72 phoneme emotion models are trained [13]. In case of emotion challenge we have 72 phoneme emotion models for two emotional classes evaluation and 180 phoneme emotion models for five emotional class.

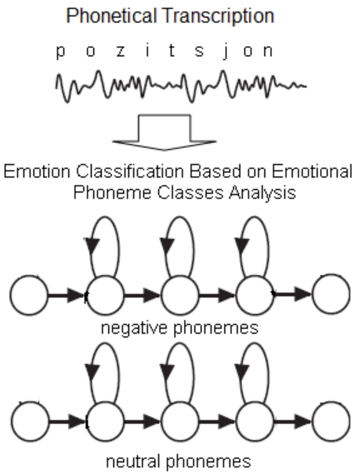

Figure 3: *Phoneme level emotion recognition.*

Emotional phonemes are modeled by training three emitting state HMM models. An HMM of 3 states and 16 Mixtures of Gaussians was built for each phoneme emotion. Standard techniques such as forward-backward and Baum-Welch reestimation algorithms were used to build these models.

In the case of phoneme level emotion analysis we can restate equation 1 in another way:

Ω is a possible emotional word (emotional phones sequence) from defined vocabulary,

$P(X|\Omega)$ is an emotion acoustic model for word Ω,

$P(\Omega)$ is the affected speech language model.

After this we are generating possible emotional phonetic transcriptions for sensible utterances by using an emotional phoneme set. In our case two transcriptions for neutral and anger speaking styles are generated, see Figure 3. After this we are using the EM algorithm to choose the most appropriate emotional transcription for the recognized sentence.

4. Experiments

All experiments are carried out in Speaker Independent and Inter-Corpora manner. In the case of evaluation of affected speech adapted HMM models we provide 10 different evaluations. For each speaker from the EMO-DB database we used 9 others for adaptation. Result values presented in Table 1 are average accuracy of speech recognition for among all evaluations. For emotional models tuning and feature set selection within emotion challenge we used LOSO strategy within training material. Average recall and accuracy for 26 independent evaluations (26 speakers in training set) were the primary measure for optimization.

4.1. Results on Affected Speech Recognition

For the first evaluation, we tested monophone HMMs trained on Kiel Corpus. Second, we evaluated monophones HMMs trained on EMO-DB Corpora in LOSO manner. Finally, we tested HMMs trained on Kiel Corpus and adapted by EMO-DB in LOSO manner.

Test corpus	Kiel	EMO-DB	Kiel adap. EMO-DB
EMO-DB[%]	85.7	92.7	97.8
Kiel[%]	93.2	89.9	93.4

Table 1: Accuracies of speech recognition for HMM models trained on Kiel corpus, EMO-DB (LOSO), Kiel models adapted by EMO-DB(LOSO), on EMO-DB and Kiel databases.

As one can see from the Table 1, the ASR engine trained on neutral samples, is not able to recognize affected speech with high accuracy just 85.7%. At the same time insufficient amount of speech data in EMO-DB is enough to provide better accuracy of emotional speech recognition about 92.7%. Finally, when we apply MAP adaptation for monophones HMM models trained on Kiel corpus based on EMO-DB sample, we get 12.1% absolute improvement of speech recognition accuracy from 85.7% to 97.8%. While, monophones HMM models adapted on affected speech provide the same recognition performance on emotionally uncolored speech from Kiel corpus.

4.2. Results on Acoustic Emotion Classification

Test-runs on EMO-DB database turn- and phoneme-level models are carried out in Leave-One-Speaker-Out (LOSO) manner to address speaker independence (SI), as required by most applications.

With turn level analysis we achieved recognition rate up to 99.0% for two classes emotion (anger and neutral) classification task and 97.3% for phone level analysis. We find out that phoneme level emotion analysis provides almost the same accuracy as turn level analysis. At the same time it provides a higher level of system integration flexibility for SDS. Both results slightly outperform results of human perception tests for two emotional classes subset 96.2%.

4.3. Results on Acoustic Emotion Classification within Emotion Challenge

As classes are substantially unbalanced, the primary measure to optimize is unweighted average (UA) recall, and secondly the weighted average (WA) recall (i. e. accuracy). Taking into account results of tuning and feature set optimization of our emotion classifier based on LOSO evaluation on training samples. We tested 25 evaluation results based on optimal configurations. The best results within challenge test set are presented below.

The best results for two classes was achieved with utterance level analysis with feature set included 12 MFCC coefficients normalized with gender dependent VTLN, Energy and their deltas and acceleration. For five classes emotion recognition task the best results were received with 13 MFCC coefficients normalized by gender dependent VTLN after CMS

Level of analysis	**Classes** [#]	**Recall** [%] **UA**	**Recall** [%] **WA**
Utterance	2	69.21	70.36
Phonemes	2	68.09	73.26
Combined	2	68.45	70.35
Baseline	2	67.7	65.5
Utterance	5	41.40	47.44
Phonemes	5	35.21	52.78
Combined	5	40.62	49.38
Baseline	5	38.2	39.2

Table 2: Accuracies and recall for emotion recognition on test set of FAU AIBO database.

included zero coefficient instead of Energy and their delta and acceleration. Confusion matrices for the best two-class and five-class presented below in Tables 3 and 4.

	NEG[#]	**IDL**[#]	All [#]
NEG	1635	830	2465
IDL	1617	4175	5792
[%]	**66.3**	**72.1**	**70.4**

Table 3: Confusion matrix for the two-classes emotion recognition task and accuracies for each class individually and all test set.

As one can see from the Table 3. for NEG class false acceptance error is quite high.

	A[#]	**E**[#]	**N**[#]	**P**[#]	**R**[#]	**All**[#]
A	315	189	67	9	31	611
E	202	944	276	10	76	1508
N	592	1551	2485	217	532	5377
P	17	17	90	53	38	215
R	95	108	176	47	120	546
[%]	**51.6**	**62.6**	**46.2**	**24.2**	**22.0**	**47.4**

Table 4: Confusion matrix for the five-classes emotion recognition task and accuracies for each class individually and all test set. Abbreviations : **A** anger, **E** emphatic, **N** neutral, **P** positive, **R** rest

In case of five emotion classes evaluation, classes are unbalanced in training set. As results we have to be very careful with over tuning of sparse emotional classes like **P, R**. At the same time the leaders of the emotion classification, see Table 4, **A , E** classes have extremely high confusions with **N** class. Providing more samples for **P, R** classes within training set can significantly increase performance for the five classes emotion classifier.

5. Conclusion

Emotions play a central role in human machine communication. We presented applicable results of emotion classification on acted emotional samples from EMO-DB within NIMITEK demonstration prototype. Also we present evaluation results comparable with baseline results, for emotion challenge based on spontaneous emotional corpus. We find out that affected speech adapted ASR engine provides higher performance of emotional speech recognition. The research simulation in this project helps to provide a close to natural way of human machine interaction. In most cases, the emotion adaptive dialog management becomes more friendly and helped the user to solve the task faster as in dialogs without emotion adaptive management model was considered.

6. ACKNOWLEDGMENTS

The work has been conducted in the framework of the NIMITEK project (http://wwwai.cs.uni-magdeburg.de/nimitek), within the framework of the Excellence Program Neurowissenschaften of the federal state of Sachsen-Anhalt, Germany (FKZ: XN3621A/1005M). This project is associated and supported by the Transregional Collaborative Research Centre SFB/TRR 62 ”Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG).

7. References

[1] C.-H. Lee, “Fundamentals and technical challenges in automatic speech recognition,” in *SPECOM2007*, Moscow, Russia, 2008, pp. 25–44.

[2] M. Gnjatović and D. Rösner, “Adaptive dialogue management in the nimitek prototype system,” in *4th IEEE PIT’08*, Kloster Irsee, Germany, 2008, pp. 14–25.

[3] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, “Towards more reality in the recognition of emotional speech,” in *IEEE ICASSP 2007*, vol. 2, Honolulu, Hawaii, USA, April 2007, pp. 941–944.

[4] C. Busso, S. Lee, and S. Narayanan, “Using neutral speech models for emotional speech analysis.” in *Interspeech 2007*, Antwerp, Belgium, 2007, pp. 2225–2228.

[5] M. Pantic and L. Rothkrantz, “Toward an affect-sensitive multimodal human-computer interaction,” in *Proccedings of the IEEE, vol. 91*, September 2003, pp. 1370–1390.

[6] A. Wendemuth, J. Braun, B. Michaelis, F. Ohl, D. Rösner, H. Scheich, and R. Warnemünde, “Neurobiologically inspired, multimodal intention recognition for technical communication systems (NIMITEK),” in *4th IEEE PIT’08*, Kloster Irsee, Germany, 2008.

[7] B. Vlasenko and A. Wendemuth, “Heading toward to the natural-way of human-machine interaction: The nimitek project.” in *IEEE ICME 2009*, Cancun, Mexico, 2009.

[8] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech,” *Interspeech*, pp. 1517–1520, 2005.

[9] B. Schuller, S. Steidl, and A. Batliner, “The interspeech 2009 emotion challenge,” in *INTERSPEECH 2009*, Brighton, United Kingdom, 2009.

[10] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University Engineering Department, 2002.

[11] B. Schuller, B. Vlasenko, D. Arsic, G. Rigoll, and A. Wendemuth, “Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition.” in *IEEE ICME 2008*, Hannover, Germany, 2008, pp. 1333–1336.

[12] P. Boersma and D. Weenink, “Praat: doing phonetics by computer (version 4.3.14) [computer program].” http://www.praat.org/, 2005.

[13] B. Vlasenko, B. Schuller, K. Tadesse, G. Rigoll, and A. Wendemuth, “Balancing spoken content adaptation and unit length in the recognition of emotion and interest,” in *INTERSPEECH 2008 incorporating SST 2008, September 23-26*, Brisbane, Australia, 2008, pp. 805–808.