# Applications of HMMs for the Recognition of Emotional Sequences in the Valence-Arousal Space

David Hübner

Otto-von-Guericke-Universität Magdeburg,
Fakultät für Elektrotechnik und Informationstechnik,
Institut für Elektronik, Signalverarbeitung und Kommunikationstechnik,
Lehrstuhl für Kognitive Systeme

20. ESSV, Dresden 2009

# Outline

1. Introduction
2. Corpus & Data Preparation
3. Hidden Markov Models
4. Validation
5. The Valence-Arousal Space
6. Conclusion & Outlook

# Outline

# Outline

# Outline

# Outline

# Outline

# Introduction

- Our aim: Simplify the man-machine interaction, so that the machine adapts in an individual way to users.
- Important aspects: Communication via speech, being aware of the users situation, emotional-mood, skills, aims...
- Keeping the dialogue short and effective.
- In this talk: How to measure a users emotions only on basis of speech in two dimensions.
- Why emotions? They have strong influence on our perception and decision making progress and different emotions may lead to different dialog strategies, e.g. by providing more/less help.

# Introduction

- Our aim: Simplify the man-machine interaction, so that the machine adapts in an individual way to users.
- Important aspects: Communication via speech, being aware of the users situation, emotional-mood, skills, aims...
- Keeping the dialogue short and effective.
- In this talk: How to measure a users emotions only on basis of speech in two dimensions.
- Why emotions? They have strong influence on our perception and decision making progress and different emotions may lead to different dialog strategies, e.g. by providing more/less help.

# Introduction

- Our aim: Simplify the man-machine interaction, so that the machine adapts in an individual way to users.
- Important aspects: Communication via speech, being aware of the users situation, emotional-mood, skills, aims...
- Keeping the dialogue short and effective.
- In this talk: How to measure a users emotions only on basis of speech in two dimensions.
- Why emotions? They have strong influence on our perception and decision making progress and different emotions may lead to different dialog strategies, e.g. by providing more/less help.

# Introduction

- Our aim: Simplify the man-machine interaction, so that the machine adapts in an individual way to users.
- Important aspects: Communication via speech, being aware of the users situation, emotional-mood, skills, aims...
- Keeping the dialogue short and effective.
- In this talk: How to measure a users emotions only on basis of speech in two dimensions.
- Why emotions? They have strong influence on our perception and decision making progress and different emotions may lead to different dialog strategies, e.g. by providing more/less help.

# Introduction

- Our aim: Simplify the man-machine interaction, so that the machine adapts in an individual way to users.
- Important aspects: Communication via speech, being aware of the users situation, emotional-mood, skills, aims...
- Keeping the dialogue short and effective.
- In this talk: How to measure a users emotions only on basis of speech in two dimensions.
- Why emotions? They have strong influence on our perception and decision making progress and different emotions may lead to different dialog strategies, e.g. by providing more/less help.

# Corpus

- Database: SmartKom Corpus (non-acted data)
- 1887 files containing up to 2 emotion transitions, e.g.
  - weak anger
  - neutral-strong joy
  - neutral-helplessness-pondering
- 1124 female and 763 male recordings.

| Original classes | New class |
|---|---|
| strong/weak joy/gratitude + strong/weak surprise | joy |
| strong/weak pondering + strong/weak helplessness | helplessness |
| strong/weak anger/irritation | anger |
| neutral + unidentifiable | neutral |

Figure: The 4 final classes for our models

# Corpus

- Database: SmartKom Corpus (non-acted data)
- 1887 files containing up to 2 emotion transitions, e.g.
  - weak anger
  - neutral-strong joy
  - neutral-helplessness-pondering
- 1124 female and 763 male recordings.

| Original classes | New class |
|---|---|
| strong/weak joy/gratitude + strong/weak surprise | joy |
| strong/weak pondering + strong/weak helplessness | helplessness |
| strong/weak anger/irritation | anger |
| neutral + unidentifiable | neutral |

Figure: The 4 final classes for our models

# Corpus

- Database: SmartKom Corpus (non-acted data)
- 1887 files containing up to 2 emotion transitions, e.g.
  - weak anger
  - neutral-strong joy
  - neutral-helplessness-pondering
- 1124 female and 763 male recordings.

| Original classes | New class |
|---|---|
| strong/weak joy/gratitude + strong/weak surprise | joy |
| strong/weak pondering + strong/weak helplessness | helplessness |
| strong/weak anger/irritation | anger |
| neutral + unidentifiable | neutral |

Figure: The 4 final classes for our models

# Distribution of the pairwise transitions in the SmartKom material

- Most frequent sequences
  - neutral-helplessness-neutral
  - neutral-joy-neutral
  - helplessness-neutral-helplessness
  - neutral-anger-neutral

| | neutral | anger | joy | helplessness |
|---|---|---|---|---|
| **neutral** | 0 | 53 | 108 | 207 |
| **anger** | 46 | 0 | 9 | 11 |
| **joy** | 62 | 2 | 0 | 15 |
| **helplessness** | 330 | 8 | 16 | 0 |

Figure: Number of pairwise transitions between the emotions

# Data Preparation

- Training difficult due to noise and long parts of silence.
- Solution: Removing the silence/noise
  - Cut out sequences with energy-values below 0.42.
  - Static threshold problematic for files with low volume (complete signal below the threshold). In this case the file remains as it is.
- Result: Enormous gain in recognition performance (up to 25%).

# Data Preparation

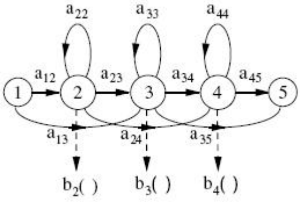
- Training difficult due to noise and long parts of silence.
- Solution: Removing the silence/noise
  - Cut out sequences with energy-values below 0.42.
  - Static threshold problematic for files with low volume (complete signal below the threshold). In this case the file remains as it is.
- Result: Enormous gain in recognition performance (up to 25%).

# Data Preparation

- Training difficult due to noise and long parts of silence.
- Solution: Removing the silence/noise
  - Cut out sequences with energy-values below 0.42.
  - Static threshold problematic for files with low volume (complete signal below the threshold). In this case the file remains as it is.
- Result: Enormous gain in recognition performance (up to 25%).

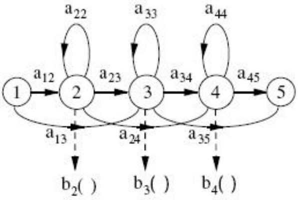
# Hidden Markov Models (HMMs) General

- Double stochastic process: 1. Dynamic of the state transitions and 2. the generation of the observable values are both stochastic.
- Postulate two-stage random process for the production of an output sequence:
  - hidden states (here:the true emotions) for the time structure in form of a Markov Chain.
  - observable outputs (here:features of the speech signal) according to a state related probability distribution.
- Successfully used in automated speech recognition.



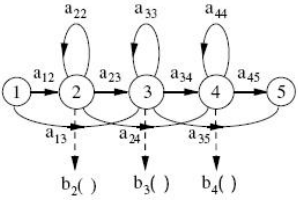
# Hidden Markov Models (HMMs) General

- Double stochastic process: 1. Dynamic of the state transitions and 2. the generation of the observable values are both stochastic.
- Postulate two-stage random process for the production of an output sequence:
  - hidden states (here:the true emotions) for the time structure in form of a Markov Chain.
  - observable outputs (here:features of the speech signal) according to a state related probability distribution.
- Successfully used in automated speech recognition.



# Hidden Markov Models (HMMs) General

- Double stochastic process: 1. Dynamic of the state transitions and 2. the generation of the observable values are both stochastic.
- Postulate two-stage random process for the production of an output sequence:
  - hidden states (here:the true emotions) for the time structure in form of a Markov Chain.
  - observable outputs (here:features of the speech signal) according to a state related probability distribution.
- Successfully used in automated speech recognition.



# The HMM Model

- **We can observe:** 39 features/sample given as the MFC-Coefficients.
- **Our aim:** Corresponding hidden sequence of emotions.
- **The Approach:** One 3-state left-right HMM/emotion, with
  - transition probabilities $a_{ij}$
  - emission probabilities for each feature $i = 1..39$ in form of a Gaussian $g_i := [\mu_i, \sigma_i]$

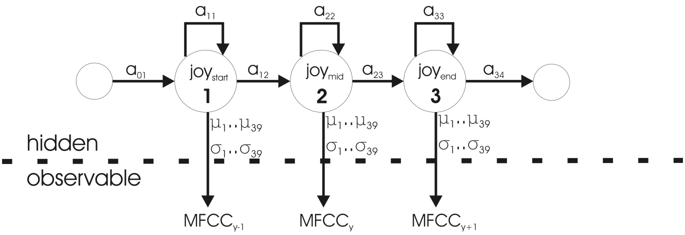



Figure: Example HMM for the emotion *joy*

# The HMM Model

- **We can observe:** 39 features/sample given as the MFC-Coefficients.
- **Our aim:** Corresponding hidden sequence of emotions.
- **The Approach:** One 3-state left-right HMM/emotion, with
  - transition probabilities $a_{ij}$
  - emission probabilities for each feature $i = 1..39$ in form of a Gaussian $g_i := [\mu_i, \sigma_i]$

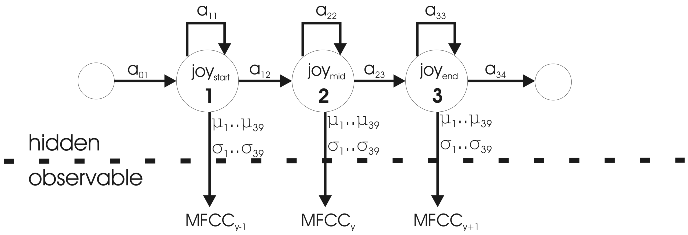



Figure: Example HMM for the emotion *joy*

# The HMM Model

- **We can observe:** 39 features/sample given as the MFC-Coefficients.
- **Our aim:** Corresponding hidden sequence of emotions.
- **The Approach:** One 3-state left-right HMM/emotion, with
  - transition probabilities $a_{ij}$
  - emission probabilities for each feature $i = 1..39$ in form of a Gaussian $g_i := [\mu_i, \sigma_i]$

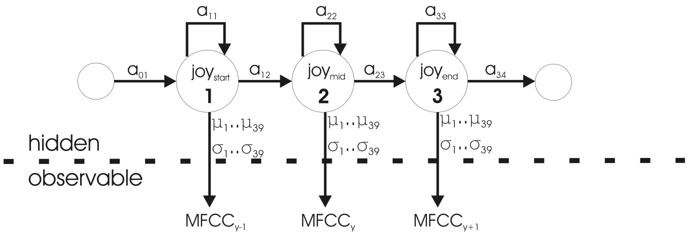



Figure: Example HMM for the emotion *joy*

# Training and Evaluation

- Training
  - Application of 5 steps of the Baum-Welch algorithm to estimate the parameters of the HMMs.
  - Why 5 steps? Best generalization!
- Evaluation with 2 Cross-Validation methods
  - 90-10: Split the data randomly in 90% training and 10% test data (50 trials).
  - Leave-One-Speaker-Out: Exclude one speaker completely from training and use him later for testing afterwards (85 female and 61 male speakers).

# Training and Evaluation

- Training
  - Application of 5 steps of the Baum-Welch algorithm to estimate the parameters of the HMMs.
  - Why 5 steps? Best generalization!
- Evaluation with 2 Cross-Validation methods
  - 90-10: Split the data randomly in 90% training and 10% test data (50 trials).
  - Leave-One-Speaker-Out: Exclude one speaker completely from training and use him later for testing afterwards (85 female and 61 male speakers).

# Results: 90-10 Cross-Validation

- The male model performs better than female one.
- The mixed model is quite close to the male one and is much better than the female model.
- The amount of training material correlates positively with recognition rate (compare neutral-joy).

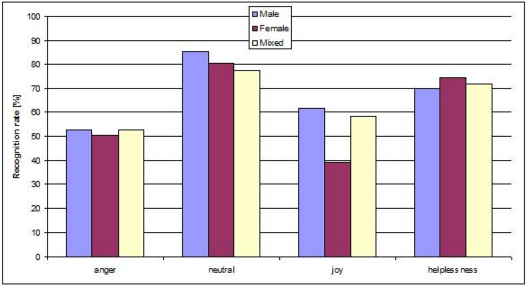



Figure: Result of the 90-10 cross validation

# Results: 90-10 Cross-Validation

- The male model performs better than female one.
- The mixed model is quite close to the male one and is much better than the female model.
- The amount of training material correlates positively with recognition rate (compare neutral-joy).

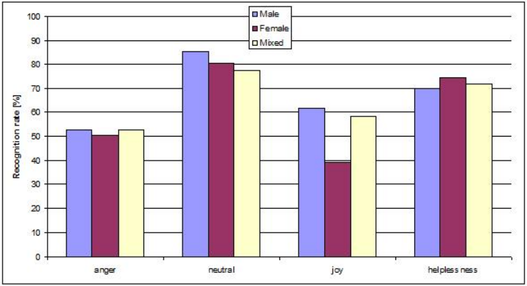



Figure: Result of the 90-10 cross validation

# Results: 90-10 Cross-Validation

- The male model performs better than female one.
- The mixed model is quite close to the male one and is much better than the female model.
- The amount of training material correlates positively with recognition rate (compare neutral-joy).

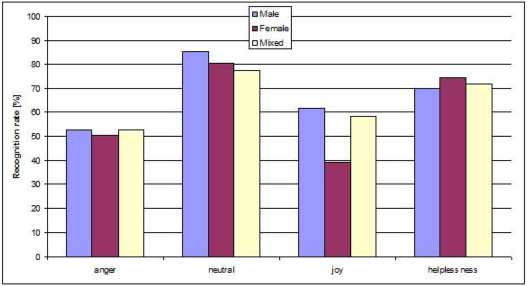



Figure: Result of the 90-10 cross validation

# Results: 90-10 vs. Leave-One-Speaker-Out Cross-Validation

- Both results are quite similar.
- Averaged over the 4 emotions the Leave-One-Speaker-Out validation performs less than 1% worse in all 3 models.
- Hence the models can be regarded as speaker independent.

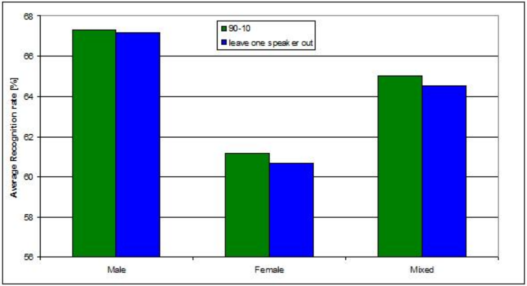



Figure: Averages of both validation methods

# Results: 90-10 vs. Leave-One-Speaker-Out Cross-Validation

- Both results are quite similar.
- Averaged over the 4 emotions the Leave-One-Speaker-Out validation performs less than 1% worse in all 3 models.
- Hence the models can be regarded as speaker independent.

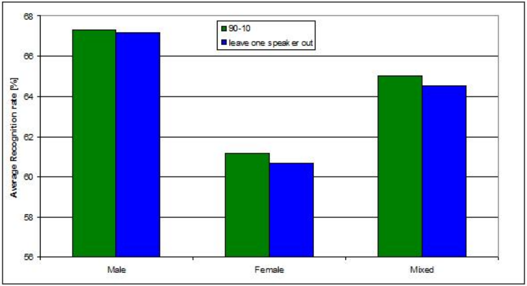



Figure: Averages of both validation methods

# Results: 90-10 vs. Leave-One-Speaker-Out Cross-Validation

- Both results are quite similar.
- Averaged over the 4 emotions the Leave-One-Speaker-Out validation performs less than 1% worse in all 3 models.
- Hence the models can be regarded as speaker independent.

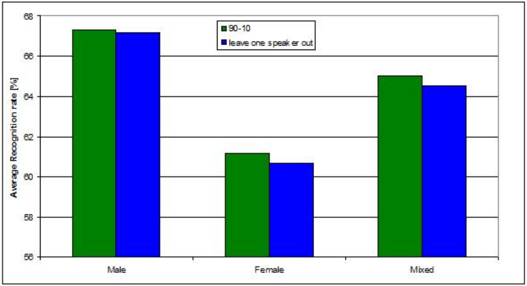



Figure: Averages of both validation methods

# The Valence-Arousal Space

- More flexible/complex than using the discrete emotion classes.
- Direct correlations between acoustic measures and the 2 dimensions (e.g. word frequency correlates with Arousal).
- First approach: Discrete mapping of the classes in SmartKom, with *helplessness* between *anger* and *neutral*.

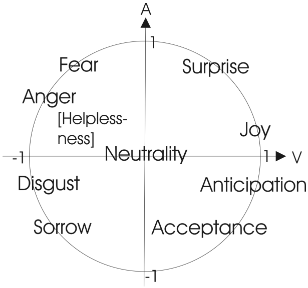



Figure: Plutchiks emotion wheel

# The Valence-Arousal Space

- More flexible/complex than using the discrete emotion classes.
- Direct correlations between acoustic measures and the 2 dimensions (e.g. word frequency correlates with Arousal).
- First approach: Discrete mapping of the classes in SmartKom, with *helplessness* between *anger* and *neutral*.

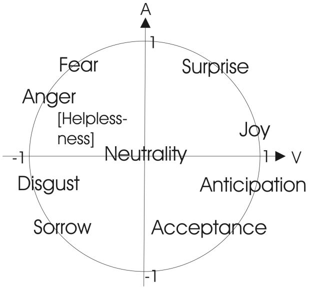



Figure: Plutchiks emotion wheel

# The Valence-Arousal Space

- More flexible/complex than using the discrete emotion classes.
- Direct correlations between acoustic measures and the 2 dimensions (e.g. word frequency correlates with Arousal).
- First approach: Discrete mapping of the classes in SmartKom, with *helplessness* between *anger* and *neutral*.

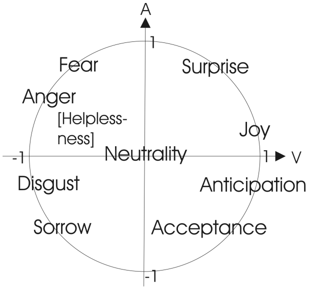



Figure: Plutchiks emotion wheel

# Determining the word frequency

- Hard task, as word boundaries not visible in the speech signal.
- Probable solution: indirect way using a [General] Speech-Recognizer.
- Here: Evaluating the transcriptions by word counting. Still 4 main problems:
  - Who is speaking?
  - Pauses between dialog turns.
  - Effects due to the removed silence.
  - Different word lengths.
- Future: Syllable frequency: maximum of the spectrum of modulation frequencies

# Determining the word frequency

- Hard task, as word boundaries not visible in the speech signal.
- Probable solution: indirect way using a [General] Speech-Recognizer.
- Here: Evaluating the transcriptions by word counting. Still 4 main problems:
    - Who is speaking?
    - Pauses between dialog turns.
    - Effects due to the removed silence.
    - Different word lengths.
- Future: Syllable frequency: maximum of the spectrum of modulation frequencies

# Determining the word frequency

- Hard task, as word boundaries not visible in the speech signal.
- Probable solution: indirect way using a [General] Speech-Recognizer.
- Here: Evaluating the transcriptions by word counting. Still 4 main problems:
  - Who is speaking?
  - Pauses between dialog turns.
  - Effects due to the removed silence.
  - Different word lengths.
- Future: Syllable frequency: maximum of the spectrum of modulation frequencies

# Determining the word frequency

- Hard task, as word boundaries not visible in the speech signal.
- Probable solution: indirect way using a [General] Speech-Recognizer.
- Here: Evaluating the transcriptions by word counting. Still 4 main problems:
  - Who is speaking?
  - Pauses between dialog turns.
  - Effects due to the removed silence.
  - Different word lengths.
- Future: Syllable frequency: maximum of the spectrum of modulation frequencies

# Mapping of our results

- Result of manual determination of 8 utterances/emotion, with minimum, maximum and average frequency.
- Frequencies between 1.5 and 4.5 were observed and used to measure the arousal A (after normalization).
- Drawback: Still static mapping of the emotions on the V-axis.

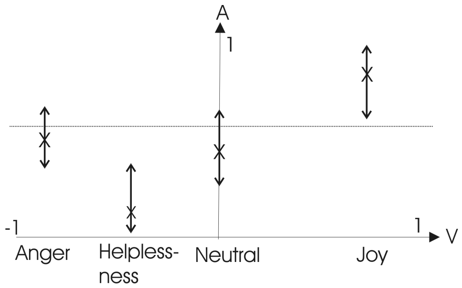



Figure: Typical word frequencies of the different emotions

# Mapping of our results

- Result of manual determination of 8 utterances/emotion, with minimum, maximum and average frequency.
- Frequencies between 1.5 and 4.5 were observed and used to measure the arousal A (after normalization).
- Drawback: Still static mapping of the emotions on the V-axis.

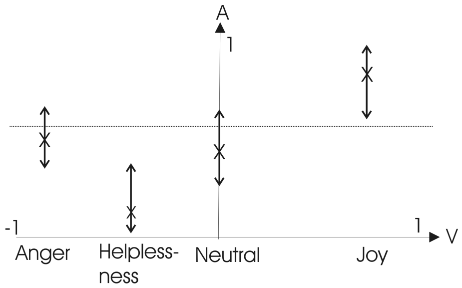



Figure: Typical word frequencies of the different emotions

# Mapping of our results

- Result of manual determination of 8 utterances/emotion, with minimum, maximum and average frequency.
- Frequencies between 1.5 and 4.5 were observed and used to measure the arousal A (after normalization).
- Drawback: Still static mapping of the emotions on the V-axis.

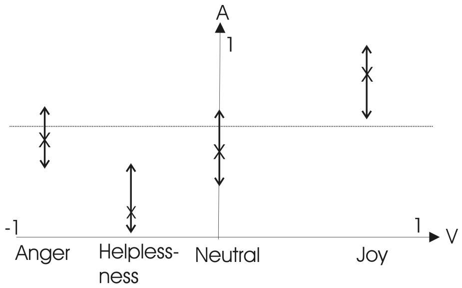



Figure: Typical word frequencies of the different emotions

# Conclusions

- Pre-processing of the training material (removing silence, noise..) pushes the recognition rates.
- Gender depend models only, iff there is enough training material available.
- A combined model may compensate the lack of material.
- The Leave-One-Speaker-Out validation proves the speaker independence of the model.
- Computing word/syllable frequencies on the basis of the speech signal is difficult.

# Conclusions

- Pre-processing of the training material (removing silence, noise..) pushes the recognition rates.
- Gender depend models only, iff there is enough training material available.
- A combined model may compensate the lack of material.
- The Leave-One-Speaker-Out validation proves the speaker independence of the model.
- Computing word/syllable frequencies on the basis of the speech signal is difficult.

# Conclusions

- Pre-processing of the training material (removing silence, noise..) pushes the recognition rates.
- Gender depend models only, iff there is enough training material available.
- A combined model may compensate the lack of material.
- The Leave-One-Speaker-Out validation proves the speaker independence of the model.
- Computing word/syllable frequencies on the basis of the speech signal is difficult.

# Conclusions

- Pre-processing of the training material (removing silence, noise..) pushes the recognition rates.
- Gender depend models only, iff there is enough training material available.
- A combined model may compensate the lack of material.
- The Leave-One-Speaker-Out validation proves the speaker independence of the model.
- Computing word/syllable frequencies on the basis of the speech signal is difficult.

# Conclusions

- Pre-processing of the training material (removing silence, noise..) pushes the recognition rates.
- Gender depend models only, iff there is enough training material available.
- A combined model may compensate the lack of material.
- The Leave-One-Speaker-Out validation proves the speaker independence of the model.
- Computing word/syllable frequencies on the basis of the speech signal is difficult.

# Outlook

- Gaining a real 2d representation, by evaluation of the provided log-likelihood's.
- Robust models for automated word/syllable determination.
- Using prior knowledge of the emotions transition probabilities and creating a history model, which helps to accept/discard the recognized emotions.

# Outlook

- Gaining a real 2d representation, by evaluation of the provided log-likelihood's.
- Robust models for automated word/syllable determination.
- Using prior knowledge of the emotions transition probabilities and creating a history model, which helps to accept/discard the recognized emotions.

# Outlook

- Gaining a real 2d representation, by evaluation of the provided log-likelihood's.
- Robust models for automated word/syllable determination.
- Using prior knowledge of the emotions transition probabilities and creating a history model, which helps to accept/discard the recognized emotions.

# References

## Thank you!

[1] SMARTKOM: Dialog-based Human-Technology Interaction Multi-Modal Database, University of Munich, 2004

[2] BATLINER, A. ; ZEISSLER, V. ; FRANK, C. ; ADELHARDT, J. ; SHI, R. ; NOETH, E. : We are not amused - but how do you know? User states in a multi-modal dialogue system. In: *Proceedings of EUROSPEECH*, 2003, S. 733 – 736

[3] PLUTCHIK, R. : The Nature of Emotions. In: *American Scientist*, 2001

[4] VOGT, T. ; ANDRE, E. : Improving Automatic Emotion Recognition from Speech via Gender Differentiation. In: *Proceedings of Language Resources and Evaluation Conference (LREC)*, 2006

[5] YOUNG, S. ; EVERMANN, G. ; GALES, M. ; HAIN, T. ; KERSHAW, D. ; LIU, X. ; MOORE, G. ; ODELL, J. ; OLLASON, D. ; POVEY, D. ; VALTCHEV, V. ; WOODLAND, P. : In: *The HTK Book (for HTK Version 3.4)*, 2006