Österreichisches Forschungsinstitut für / Austrian Research Institute for / Artificial Intelligence

TR-2007-11

Hannes Pirker

Phonetic Segmentation of the GEMEP-Corpus: Applying Forced Alignment to Emotional Speech

• Freyung 6/6 • A-1010 Vienna • Austria •
• Phone: +43-1-5336112 •
• mailto:sec@ofai.at •
• http://www.ofai.at/ •

Österreichisches Forschungsinstitut für / Austrian Research Institute for / Artificial Intelligence

TR-2007-11

Hannes Pirker

Phonetic Segmentation of the GEMEP-Corpus: Applying Forced Alignment to Emotional Speech

The Austrian Research Institute for Artificial Intelligence is supported by the Federal Ministry of Education, Science and Culture.

Citation: Pirker H.: Phonetic Segmentation of the GEMEP-Corpus: Applying Forced Alignment to Emotional Speech. Technical Report, Österreichisches Forschungsinstitut für Artificial Intelligence, Wien, TR-2007-11.

Contents

(This page was intentionally left empty)

Abstract

This report documents the efforts of applying MFCC based Hidden Markov Models for the task of phonetic segmentation of *emotional speech*.

The samples of emotional speech were taken from the *Geneva Multimodal Emotion Portrayals* (GEMEP) corpus. This multimodal corpus of acted emotional utterances provides data with highly uniform and controlled lexical content, and thus offers a promising basis for further systematic studies, especially on the acoustic properties of emotional speech as well as on the temporal relationship between speech, gestures and facial expressions. The phonetic segmentation on the level of phonemes described in this report offers a solid basis for all kinds of further investigations of temporal properties of multimodal emotional data.

The report provides a description of the technical lay-out of the automatic alignment-procedure, observations on peculiarities of the data and an evaluation of the obtained quality of the segmentation.

1 Introduction

The primary motivation for this study lies in our interest in multimodal generation of speech and speech accompanying gestures in the context of Embodied Conversational Agents. More specifically we were in search of a more profound empirical basis for investigating the temporal synchronisation between gestures and speech. The *Geneva Multimodal Emotion Portrayals* corpus (GEMEP) [Bänziger et al. 2006] [Bänziger et al. 2007] a multimodal corpus of acted emotional utterances with highly controlled uniform content was identified as a promising basis for performing research in this area.

As we are specifically interested in the temporal aspects of multimodal interactions performing a segmentation of the audio data is a prerequisite for acquiring fine grained temporal anchor points for all further analysis.

The standard way to perform automatic segmentation in this context is to use forced alignment, which typically is implemented by building a restricted speech recogniser that uses Hidden Markov Models (HMM) for the statistical modelling on the basis of Mel Frequency Cepstral Coefficients (MFCC) for encoding the speech.

Though the technical concepts for constructing such an aligner are established as good practise, it was by no means clear beforehand on how well the methods would work in the context of emotional speech. Automatic phonetic alignment is most favorably performed on single speakers and more importantly usually deals with modal voices and typically read speech. Thus is was an open question on how the MFCC-based phonetic segmentation would deal with speech from a corpus with a high degree of variation in voice qualities and speaking modes, ranging from flustering to outright shouting.

This report documents the alignment procedure and provides an evaluation of the obtained quality. In the Appendix further information e.g. on the labelling scheme used and on observations on peculiarities found in the database is provided for possible future users of the data.

2 Database Description

2.1 Database Design

The subject of this study is the *Geneva Multimodal Emotion Portrayals* corpus (GEMEP) [Bänziger et al. 2006] [Bänziger et al. 2007], which consists of audio and video recordings of 10 professional French speaking actors, 5 of which are female.

The actors were asked to improvise an interaction with the director where a requested affective state was to be expressed. The verbal content of the interactions used for this study is restricted to only 2 different pseudo-linguistic sentences presented in Table 1:

	orthographic representaion	SAMPA representation
type1	*Né kal ibam soud molén!*	/ne kal ibam sud mo len/
type2	*Koun sé mina lod bélam?*	/kun se mina lod belam/

Table 1: *The 2 pseudo-linguistic sentences from GEMEP used for this study.*

The actors were provided with the orthographic representation only.

As can be seen from the phonetic transcription provided in SAMPA[1], these meaningless sentences – henceforth distinguished as *type1* and *type2* – display the same number of syllables and share the same set of phonemes.

The designers of the GEMEP corpus put much effort into the choice of emotions used in their recordings. Apart from including emotions most frequently studied in the literature they also aimed for a more balanced distribution between positive and negative emotions. Multiple samples from the same family of emotions that are differentiated in their level of arousal where included as well.

This results in a set of 18 emotions which are enumerated in alphabetic order in Table 2. Table 3 gives an overview of the grouping of a core set of 12 emotions in respect to their valence and activation.

Throughout the recording session the actors produces several repetitions for the requested emotion and (if applicable) intensity level. The number of recorded repetitions was not fixed but was subject to the rating of the supervising director. The mean number of repetitions for each token in the corpus is 7.9.

Altogether this design results in 3815 samples: 2739 sentences of type1 and 1076 sentences of type2.

2.2 Recording Conditions

The audio data was recorded with a sampling rate of 44100 Hz while video-taping facial expressions and body movements of the actors. A clip-on microphone positioned over the actort's left ear was used in order to avoid the occlusion of facial features in the video.

For each actor the gain level was set to a fixed value prior to the recording session so that no clipping would appear at the actor's simulated 'maximal' (shouting) volume. On the one hand this constant gain level allows for the direct comparison of signal amplitudes

[1]SAMPA computer readable phonetic alphabet.
Cf. `http://www.phon.ucl.ac.uk/home/sampa/index.html`, last visited: 10.10.2007.

Shortcut	French expression	English translation	# of samples
adm	admiration	admiration	83
amu	amusement	amusement	179
att	attendrissement	tenderizing	88
col	colère chaude	hot anger	210
deg	dégoût	disgust	83
des	désespoir	despair	242
fie	fierté	pride	316
hon	honte	shame	126
inq	inquiétude	anxiety	378
int	intérêt	interest	319
irr	irritation	irritation	270
joi	joie exaltée	(exalted) joy	215
mep	mépris	contempt	92
peu	peur panique	(panic) fear	250
pla	plaisir sensual	(sensual) pleasure	263
sou	soulagement	relief	346
sur	surprise	surprise	95
tri	tristesse	sadness	260

Table 2: *The 18 emotions displayed in GEMEP and the number of occurrences in the corpus.*

	positive		**negative**		**Additional emotions**	
high	joi (joy)	215	col (hot anger)	210	hon (shame)	126
	amu(sement)	179	peu (panic fear)	250	deg (disgust)	83
	fie (pride)	316	des(pair)	242	sur(prise)	95
low	pla (pleasure)	263	irr(itation)	270	att (sentimentality)	88
	sou (relief)	346	inq (anxiety)	378	adm(miration)	83
	int(erest)	319	tri (sadness)	260	mep (contempt)	92

Table 3: *12 primary emotions in GEMEP grouped along the dimension activation (high,low) and valence (positive, negative) plus 6 additional emotions.*

across samples. On the other hand this results in a considerable number of recordings with extremely low gain levels and thus unfavorable signal-to-noise ratios.

The recordings were not performed in a noise cancelled environment which resulted in clearly perceivable reverberations in a number of recordings.

Typical problems thus occur when the exact end of an utterance is to be determined. Due to the missing gain-adjustment the final sound of samples with very low energy may melt into the background noise. On the other hand, for samples spoken with high effort, considerable reverberations sometimes make it difficult to correctly determine the actual end of the utterance.

3 Phonetic Segmentation

The standard technique for retrieving the location of boundaries between phonemes in a speech signal is forced alignment.

Strongly simplifying, the basic idea is to use the same methods as in standard speech recognition systems. When building a speech recogniser first a manually segmented corpus is used in order to train a separate Hidden Markov Model (HMM) for each phoneme. In the recognition phase the recogniser then is searching for the sequence of phonemes that provides the most likely match with the input signal.

In the case of forced alignment the recognition phase is different insofar as the segmental content of the input signal is already known beforehand (e.g. *'Ne kal ibam...'* in our case) and the decoder now 'only' needs to determine the boundary *positions*, i.e. the most likely transitions between subsequent models.

In this section the single steps are described in more detail.

3.1 Manual Labelling

The labelled data necessary for training the HMM-models of the automatic aligner was constructed using a bootstrapping procedure: a small set of sentences is manually segmented, then a first version of the aligner is trained on this preliminary training corpus. Alignment-results from this first aligner-version are then manually corrected, providing a now larger training corpus. This data then is used in order to train an improved version of the aligner, which in turn provides new data for increasing the set of available training data etc.

All manual labelling was performed using the highly recommendable Speech Filing System (SFS)[2].

For the manual labelling auditive cues as well as visual cues from both the wave-form and a wide-band spectrogram were taken into account.

In order to get the system launched more quickly, manual labelling and training of models was first performed for sentences of type 1 exclusively. Only at a rather late stage in the development the aligner was applied to sentences of type 2 as well, and from then on both type1 and type2 data was integrated into the same bootstrapping cycles.

For transcribing the phonetic content basically SAMPA was used. SAMPA labels were enhanced with special diacritics in order to explicitly encode deviations from the canonical transcription (i.e. insertions and substitutions). Diacritics were also appended to the SAMPA labels in order to mark peculiarities in the voice quality or in order to mark specific manners of articulation. An additional set of labels for events like breathing and emotive bursts (e.g. `BURST_LAUGH`, `BURST_COUGH`) was developed in a bottom up fashion throughout the labelling process.

The resulting final labelling scheme is documented in full detail in Appendix A.

[2]SFS is provided by Mark Huckvale from University College London (UCL) and is freely available at `http://www.phon.ucl.ac.uk/resource/sfs/`. There also is a tutorial available on how to use SFS together with HTK for performing phonetic segmentation. Cf. `http://www.phon.ucl.ac.uk/resource/sfs/howto/htk.htm`. Though in our own work we relied on different procedures developed earlier at our site, readers who are interested in building their own aligner from scratch are highly recommended to consult this tutorial.

3.2 Grammar and Dictionary

For the decoding a recognition grammar that deals with variations on the segmental level observed in the data was manually constructed (Cf. Appendix E). The most important aspects are the optional insertion of pauses between words and the frequent insertion of syllable-final schwa-sounds, which are due to the influence of the speaker's native language French in this pseudo linguistic data. Type1 sentences are thus frequently pronounced as `/ne kal ibam@ _ sud molEn@q /`. The distribution of these optionally inserted `/@/`-sounds is strongly unbalanced btw. and depends on the identity of the speaker, the sentence type and the emotion!

The highly uniform segmental content in the GEMEP corpus (only two different sentences all together) allows for the training of *phonemes in context models*. Separate models are trained for one and the same phoneme when occurring in different phonetic contexts, e.g. two different models are trained for the two occurrences of `/a/` in *'Ne k***a***l ib***a***m...'*.

This was achieved by simply using separate labels for each occurrence of a phoneme in the grammar. As depicted in the sample grammar below, e.g. the both variants of `/a/` are referred to as A1 and A2 respectively. The labels in the transcription files are renamed by appending the name of its proceeding neighbour and thus transforming them to unique bigram names, i.e. `/n e k a l .../` is recoded as `/n%e e%k k%a l%.../`.

The non-terminal labels used in the grammar file are then matched to these context-dependent bigram-names in a separate dictionary file:

```
EXCERPT form GRAMMAR
--------------------
(PAUSSTART)
      N1 E1 (PAUSE)
      K A1 (L1) (SCHWA1) (PAUSE)
      I B A2 M1 (SCHWA2)
            ...
(PAUSEND)

EXCERPT from DICTIONARY:
------------------------
A1 -> (a%l, a%i, a%_)
A2 -> (a%m)
```

The full grammar and dictionaries can be found in Appendix E, F and G.

3.3 Feature Extraction

The original audio data was downsampled from 44.1 kHz to 22.05 kHz and high-pass filtered at 55 Hz. For the calculation of the MFCC a frame size of 30 ms was used. A window shift of only 2.5 ms (i.e. resulting in 400 frames per second) was chosen in order to obtain the required temporal resolution for the phonetic segmentation task. 12 MFCC, energy, delta and acceleration were included resulting in a 39 dimensional feature vector. After preliminary inconclusive comparisons of MFCC with and without energy normalisation it was decided to proceed with non-normalised MFCC.

3.4 Aligner Construction

For each phoneme a left-to-right models Hidden Markov Models with 3 states and 5 Gaussian mixtures in each state were employed. Additional models were trained for noise, breathing sounds, initial, final, and sentence-internal pauses.

Baum-Welch algorithm was used for training the HMMs and Viterbi decoding is performed in order to retrieve the segment boundaries. Extraction of MFCC features, training and application of HMMs was performed with the respective tools provided by the Hidden Markov Toolkit HTK [Young et al. 2006].

4 Evaluation of Results

4.1 Test- and Training-Sets

As a result of the reiterated application of training, alignment and manual correction finally a total of 1313 sentences with manual segmentation or manually corrected alignment results are made available.

In order to provide a simple and consistent criterion for splitting this data into test- and training-sets, the number of repetitions is used. For each recording condition (i.e. each actor, each emotion, sentence type and regulation condition) several repetitions were recorded. The counter that identifies each sample as the n^{th} repetition was chosen for identifying different sub-sets of the corpus. Most importantly, recordings with repetition counter 3 and 4 (i.e. the 3^{rd} and 4^{th} repetition), were reserved for testing only.

The corpus of overall 1313 sentences with validated segmentation was thus split into a training set with 892 sentences (`TRAIN_n892_of_n1313`) and a test set comprising 421 sentences (`TEST_n421_of_n1313`), i.e. an overall split of approximately 70% for training and 30% for testing was obtained.

In order to allow for the evaluation of the effect of the amount of data used for training, additional smaller sets of training data were created, containing 505, 225, 119 and 61 sentences.

In Appendix D the content of the different test- and training-sets used in this report is documented in more detail.

4.2 Performance Measures

For the evaluation of the aligner performance the absolute value of the difference between automatically aligned and manually corrected *initial* phoneme boundaries is used.

In this context, the usage of quantiles and error thresholds proved to be a useful and meaningful measure. They both provide a impression of the proportion of correctly identified boundary locations and those which require manual correction. The measure of both mean and median errors, on the other hand are less instructive as the distribution of the absolute error is strongly right skewed. All statistics presented here were produced using the open-source statistical package R [R Core Team 2007].

Two different error measures are presented namely segment-wise and sentence-wise evaluation.

In table 4 the percentage of segments that lie within a given error-distance is enumerated. In table 5 the *maximum* error within each sentence is analysed. This later measure

gives an impression of the proportion of whole sentences with no further need for manual correction.

% of SEGMENTS w/ Error < Threshold		
N train=892 (typ1 n=685, typ2 n=207)		
Threshold	Type1	Type2
10[ms]	74.7	66.0
20[ms]	86.4	79.2
40[ms]	93.4	90.3
60[ms]	96.1	93.2
100[ms]	97.9	95.8

Table 4: Percentage of SEGMENTS where alignment error does not exceed the given threshold in [ms].

% of SENTENCES w/ MAXIMUM Error < Threshold		
N train=892 (typ1 n=685, typ2 n=207)		
Threshold	Type1	Type2
10[ms]	4.2	0.8
20[ms]	19.8	9.9
40[ms]	47.3	39.7
60[ms]	65.0	52.7
100[ms]	82.3	70.2

Table 5: Percentage of SENTENCES where the *maximal* alignment error within the sentence does not not exceed the given threshold in [ms]

In both tables, results for the two sentence-types are provided separately. The number of training-samples for type 1 is more than 3 times higher than for type 2, so the superior results for type 1 are not surprising.

If one assumes that an segmentation error of 20ms is to be a tolerable deviation, it can be seen, that approx. 80%-85% of phoneme-boundaries are located within this limit. The analysis of results on the sentence-level on the other hand display, that only 10%-20% of sentences would not need any further corrections, as the maximum error to occur in the whole sentence would be less than 20ms.

4.3 Influence of Amount of Training-Data

As described above the whole project of phonetic segmentation was performed using a bootstrapping procedure. The aligner models were trained with ever increasing number

of training data. In this section it is analysed how the performance of the aligner behave in respect to the amount of training-data used[3].

Table 6 and 7 enumerate the performance measures for aligners which were trained on 61, 119, 225, 505 and 892 sentences and were tested on the standard test-corpus with 421 sentences (Cf. Appendix D for a detailed description of these sub-corpora).

Alignment Evaluation: % of SEGMENTS w/ Error < Threshold										
Size of Train-Set	**N=61** (typ1 n=46, typ2 n=15)		**N=119** (typ1 n=87, typ2 n=32)		**N=225** (typ1 n=177, typ2 n=48)		**N=505** (typ1 n=395, typ2 n=108)		**N=892** (typ1 n=685, typ2 n=207)	
Method	Global		Global		Global		Global		Global	
Threshold	typ1	typ2	typ1	typ2	typ1	typ2	typ1	typ2	typ1	typ2
10[ms]	62.2	47.9	67.0	55.5	72.2	58.1	74.3	63.8	74.7	66.0
20[ms]	75.9	60.9	79.6	71.5	83.8	74.9	85.6	77.6	86.4	79.2
40[ms]	86.2	73.3	88.7	84.4	91.7	85.7	93.0	89.2	93.4	90.3
60[ms]	90.2	79.0	92.4	88.9	94.6	90.2	95.6	92.4	96.1	93.2
100[ms]	93.5	87.0	95.1	93.0	96.7	93.8	97.8	95.6	97.9	95.8

Table 6: Comparing performance with differently sized training sets. Percentage of segments where alignment error does not exceed the given threshold in [ms]

Alignment Evaluation: % of SENTENCES w/ MAXIMUM Error < Threshold										
Size of Train-Set	**N=61** (typ1 n=46, typ2 n=15)		**N=119** (typ1 n=87, typ2 n=32)		**N=225** (typ1 n=177, typ2 n=48)		**N=505** (typ1 n=395, typ2 n=108)		**N=892** (typ1 n=685, typ2 n=207)	
Method	Global		Global		Global		Global		Global	
Threshold	typ1	typ2	typ1	typ2	typ1	typ2	typ1	typ2	typ1	typ2
10[ms]	0.7	0.0	0.7	0.0	2.1	0.0	4.3	0.8	4.2	0.8
20[ms]	8.8	0.0	9.8	0.8	15.7	6.1	19.5	5.4	19.8	9.9
40[ms]	24.3	4.5	34.4	20.3	43.4	28.0	46.8	27.7	47.3	39.7
60[ms]	39.1	9.8	50.5	42.1	59.4	43.2	61.3	50.8	65.0	52.7
100[ms]	58.1	27.8	66.3	62.4	73.4	59.8	80.9	69.2	82.3	70.2

Table 7: Comparing performance with differently sized training sets. Percentage of SENTENCES where the *maximal* alignment error within the sentence does not not exceed the given threshold in [ms]

In Figure 1 portions of the same data are presented grafically for easier evaluation.

[3]One of the problems of evaluating forced alignment is, that not only training-data has to be tediously provided, but it also takes some work to come up with manually segmented data to be usable for testing the actual quality. Therefore most of the evaluations provided in this section only became possible *after* the whole bootstrapping-procedure was finished and a reasonably amount of data became available to be used as test-data.

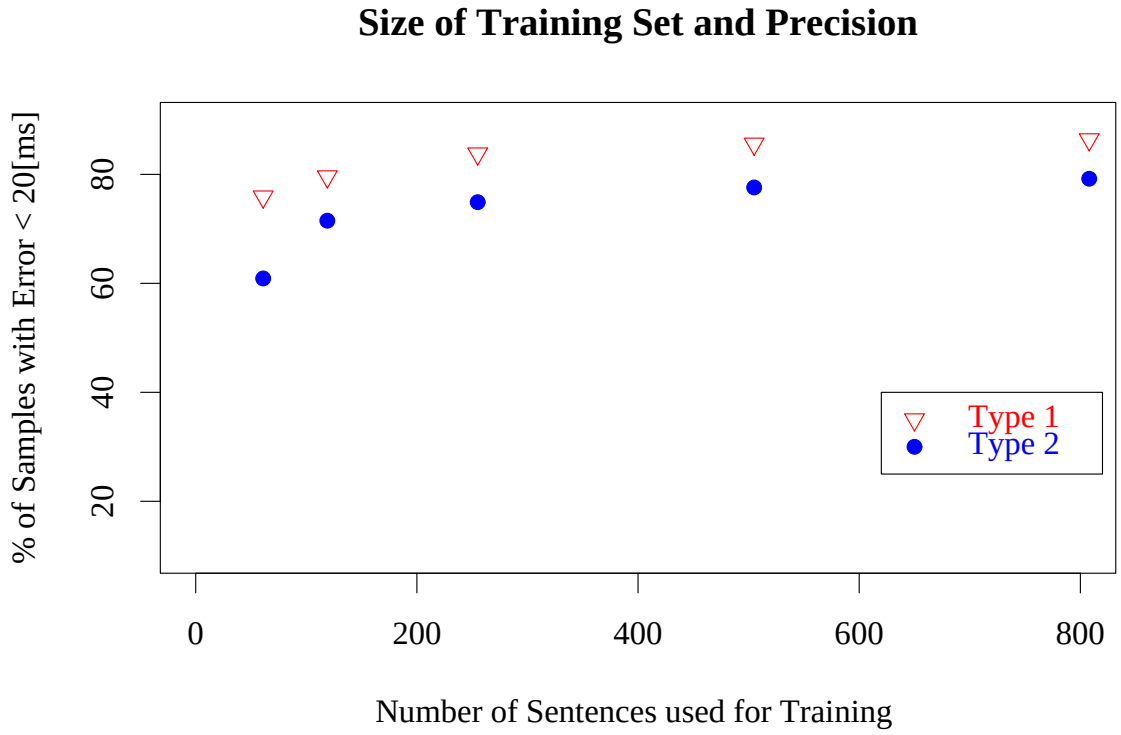


Figure 1: Achieved precision of the alignment in dependency of the amount of training data: Comparing results from training-sets with 61,119,225, 505 and 892 sentences.

The results from this objective evaluation are in line with the subjective experience throughout the bootstrapping-process: the alignment procedure is able to come up with reasonable results quite soon, i.e. with a relatively small set training data. On the other hand the results soon hit a ceiling: increasing the number of training samples only brings small improvements in the alignment results.

The explanation from subjective experiences is, that the phoneme models are able to correctly identified the majority of transitions between sounds with a very limited amount of training data. On the other hand approx. 15% of samples 'stubbornly' remain problematic, even the amount of training data is considerably increased.

4.4 Different Training Designs

In order to evaluate the respective performance of speaker-independent, speaker-dependent gender-dependent and emotion-dependent models, four different designs were employed for building additional variants of the aligner.

1. `Global`: The standard model. Use the whole training set for building speaker independent models, i.e. build one model per phoneme-in-context.

2. `perSpeaker`: Train separate, fully speaker specific models for each actor, i.e. build 10 models per phoneme.

3. `perGender`: Train models separately for male and female speakers, i.e. build 2 models per phoneme.

4. `perEmo`: Train separate models for each emotion, i.e. build 18 models per phoneme.

The results of the different training methods can be compared in Table 8 and Table 9. They show a slight advantage of the `perGender` models over the standard `Global` model. On the other hand the differences between the models are remarkably small, given that the training data available for each phoneme-model varies greatly, i.e. for the `Global` model in average there are 18 times more samples available for training a single HMM than for the `perEmo` models.

Alignment Evaluation: % of SEGMENTS w/ Error < Threshold								
Size of Train-Set	**N train=892** (typ1 n=685, typ2 n=207)							
Method	**Global**		perGender		perSpeaker		perEmo	
Threshold	typ1	typ2	typ1	typ2	typ1	typ2	typ1	typ2
10[ms]	74.7	66.0	74.5	64.3	72.0	57.2	71.9	58.0
20[ms]	86.4	79.2	87.3	78.9	84.8	73.4	85.1	73.6
40[ms]	93.4	90.3	94.3	89.1	92.9	84.8	93.7	86.0
60[ms]	96.1	93.2	96.4	93.0	95.4	89.2	96.4	90.7
100[ms]	97.9	95.8	98.2	95.5	97.6	93.9	98.0	94.3

Table 8: Percentage of SEGMENTS where alignment error does not exceed the given threshold in [ms]. Comparison of different training modes.

Alignment Evaluation: % of SENTENCES w/ MAXIMUM Error < Threshold								
Size of Training-Set	**N train=892** (typ1 n=685, typ2 n=207)							
Method	**Global**		perGender		perSpeaker		perEmo	
Threshold	typ1	typ2	typ1	typ2	typ1	typ2	typ1	typ2
10[ms]	4.2	0.8	4.2	0.8	5.3	0.8	3.8	0.8
20[ms]	19.8	9.9	23.9	8.4	20.4	6.8	18.2	4.6
40[ms]	47.3	39.7	53.7	29.0	50.2	23.3	52.1	21.4
60[ms]	65.0	52.7	66.7	53.4	64.9	36.8	71.3	42.0
100[ms]	82.3	70.2	83.2	70.2	79.6	60.2	84.3	59.5

Table 9: Percentage of SENTENCES where the *maximal* alignment error within the sentence does not exceed the given threshold in [ms]. Comparison of different training modes.

These results again indicate that the amount of training data is not the main issue. Nice segmentations can be obtained for the majority of transitions no matter which method is used, but it is very hard to further improve on the set of problematic cases.

This also holds true for the usage of different MFCC-feature sets: in addition to the MFCC features without energy normalisation used throughout this study, models using MFCC *with* energy normalisation were trained. Table 10 shows that energy normalisation does perform slightly better than the non-normalised MFCCs, but again the differences are surprisingly small.

Comparison of different MFCC-models				
% of SEGMENTS **w/ Error < Threshold**				
N train=892 (typ1 n=685, typ2 n=207)				
Energy Normalisation:	NO		**YES**	
Threshold	Type1	Type2	Type1	Type2
10[ms]	74.70	66.00	75.40	65.00
20[ms]	86.40	79.20	86.90	80.00
40[ms]	93.40	90.30	93.90	90.90
60[ms]	96.10	93.20	96.00	94.20
100[ms]	97.90	95.80	97.60	96.70

Table 10: Percentage of SEGMENTS where alignment error does not exceed the given threshold in [ms]. Comparing MFCC feature-vectors with and without energy normalisation.

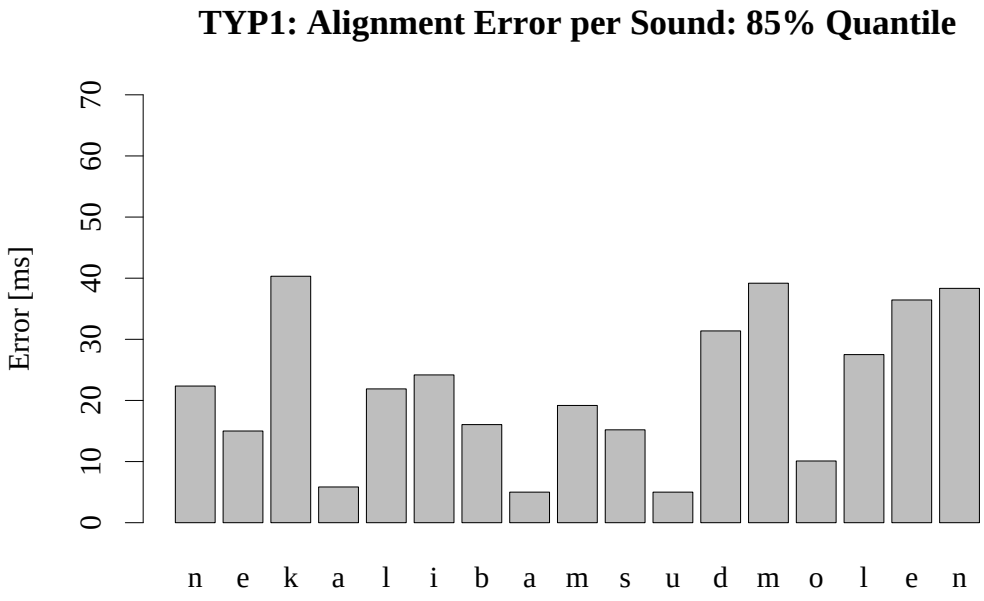
4.5 Differences per Phoneme, Actor, Emotion

The presentation of the alignment results is to be concluded with some additional analysis of the precision achieved on different sub-sets of the data.

Figure 2 depicts the achieved precision for the different sounds in the two sentences. Unsurprisingly the boundaries of the only fricative sound `/s/` are nicely located. Also the boundary detection between the plosive `/b/` and the proceeding vowel is very reliable. The detection of the start of the plosive itself is much more problematic[4]

Both subjective impression and empirical evaluation indicate that, unsurprisingly, the identity of the speaker as well as the expressed emotion influence the quality of the alignment. The respective results are presented in Fig. 3 and Fig. 4.

[4] That the worst performance is found in the transitions between `/d/` and`/b/` in the *lod belam*-examples can be explained as follows: In many of the samples`/d/` and`/b/` did not leave any clear impression in the recordings but merged into a single closure pause. When manually labelling the data, in absence of other cues the boundary between`/d/` and`/b/` was by convention just placed in the middle of the closure pause. The HMM-models then could not be blamed for not correctly differentiate between two stretches of virtual silence.



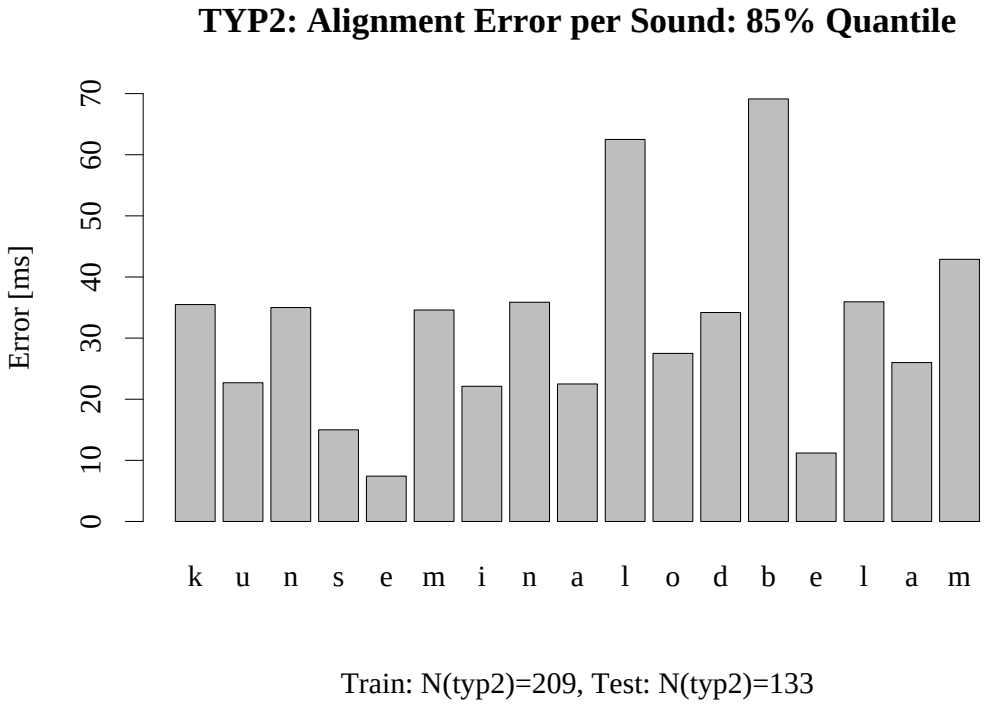
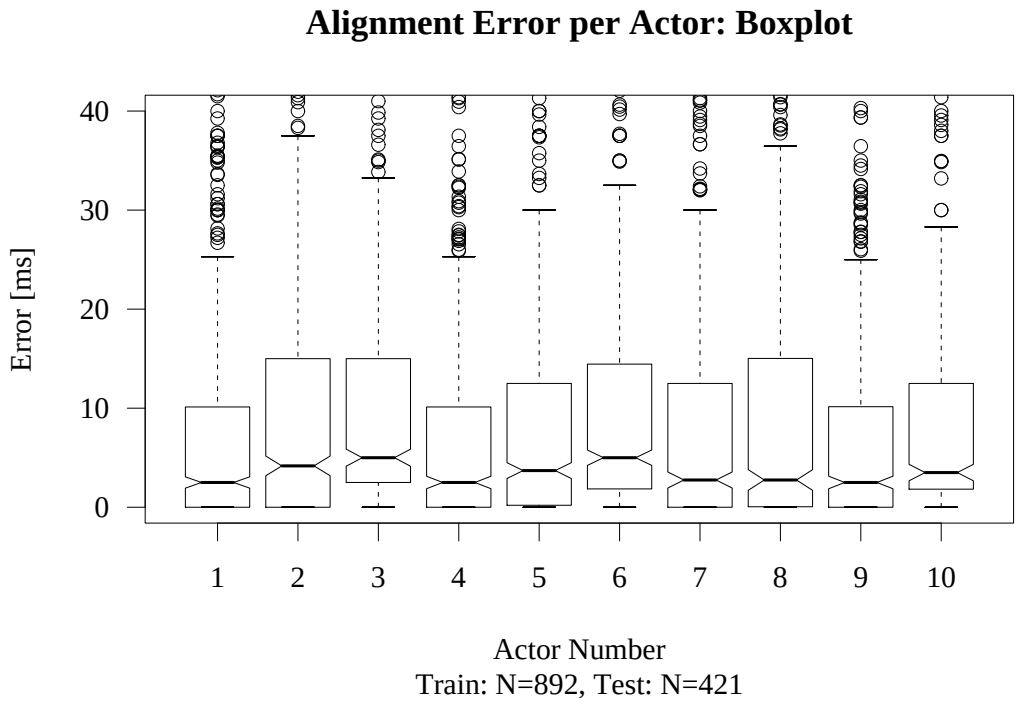


Figure 2: Alignment-error [ms] per segment: 85% quantile for each segment in sentence of type1 and type2



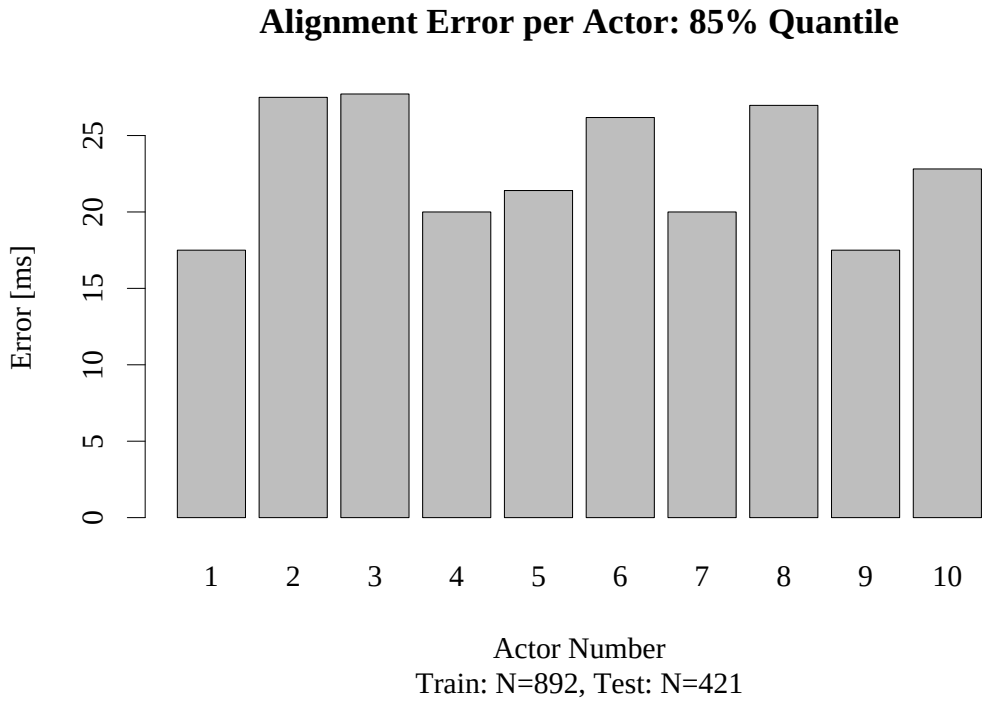


Figure 3: Alignment-error [ms] per actor: Boxplot and 85% quantile

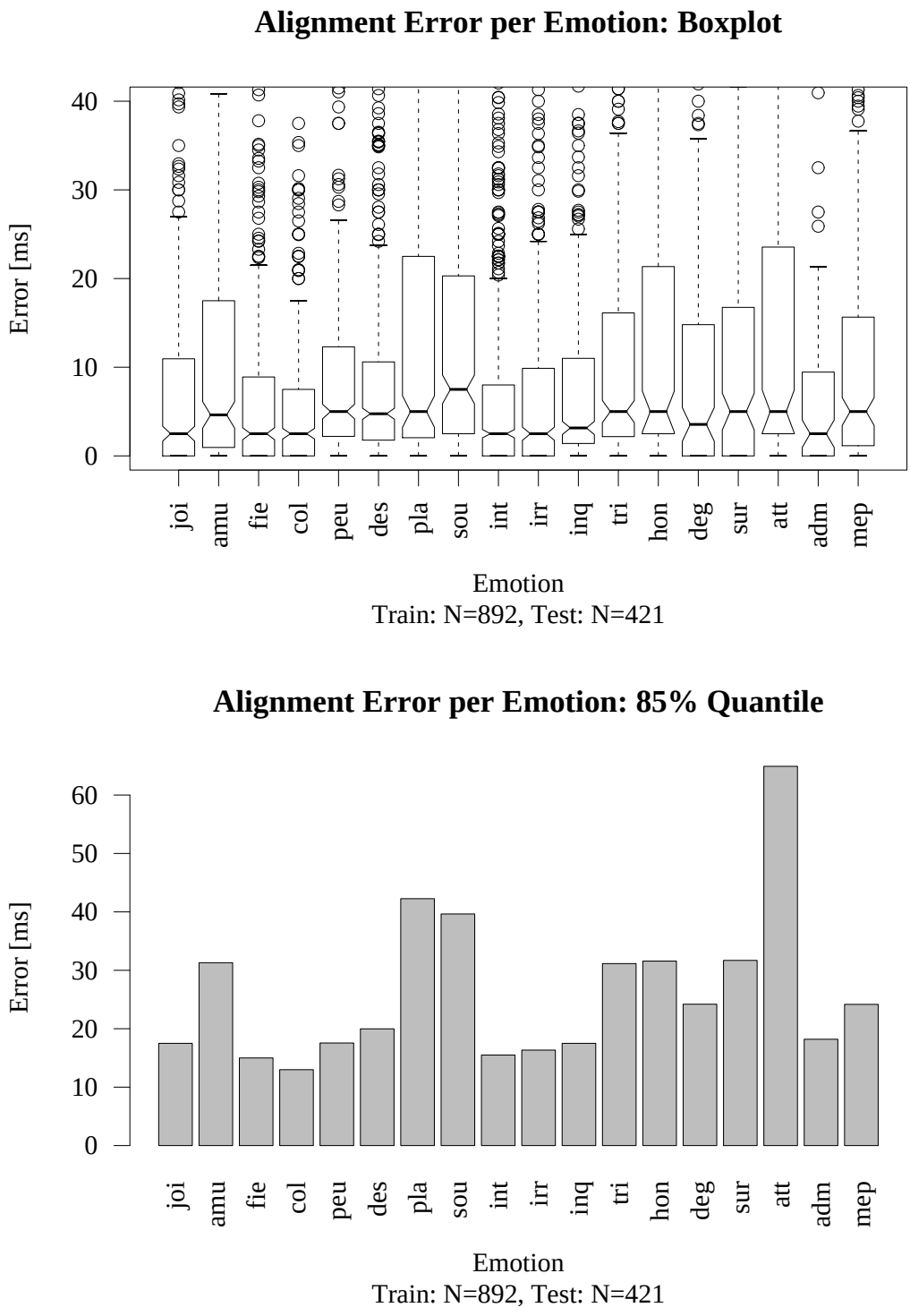


Figure 4: Alignment-error [ms] per emotion: Boxplot and 85% quantile

4.6 Discussion

When comparing the results obtained with the varying training designs, differently sized training sets and MFCC feature vectors it becomes clear that though there do exist differences in the obtained quality, these differences are relatively modest. Of course results do improve when the amount of training data is increased, but as Fig. 1 clearly depicts, these improvements soon become gradually.

For the comparison of different training designs we conclude, that though the gender specific `perGender` model does perform slightly better than the standard model it is probably not worth taking the additional effort, most specifically at the very beginning of the bootstrapping process. From a practical point of view the most relevant lesson learnt is, that the organisationally most simple `Global` procedure does perform reasonably well and that especially speaker dependent and emotion dependent models are not worth the effort.

In all evaluations the results for type2 sentences are considerably worse than for type1. On the one hand this can be explained by the differences in the amount of training data: all training sets used contain approximately 3 times more samples of type1 than of type2. On the other hand, there remain differences in performance even when this quantity-effect is taken into account. This can attributed to a phonetic structure less felicitous for segmentation. E.g. the exact start of the initial plosive `/k/` as well as the boundary between the adjacent plosives `/d/` and `/b/` are an especially tricky problem even for the human labeler.

In phonetic terms the variations of the performance can be attributed to differences in voice quality, precision of articulation and deviations from the canonical segmental content. A most critical case is, e.g., flustered speech as it both gets an unfavorable signal-to-noise ratio and weakly articulated formant structures. It often is also combined with sloppy articulation, resulting in weak or missing occlusions in stop sounds or even in the omission of whole segments. Soft speech even of very low intensity on the other hand seems less problematic. The same holds for samples of loud or even shouted speech. They are often combined with clear articulation or even over-articulation which facilitates the segmentation process.

Abnormalities in the segmental content, e.g. stuttering, are problematic but infrequent. A distinctively problematic case frequently found, e.g. in samples of amused speech, is the presence of laughter within the utterance, which poses severe problems to every speech recogniser or aligner.

The differences in the performance of speakers can be explained by just these factors. On the one hand there are natural differences in the clarity of spectra produced by the speakers. On the other hand the differences can be attributed to the varying expressive means these actors choose for encoding the intended emotions. Speakers that choose extreme speaking modes and voice qualities or frequently insert laughter and other emotive bursts in their utterances obviously pose more difficulties for the alignment procedure.

The unequal performance of the different emotions also reflects the observations mentioned above. On the one hand emotions typically produced with a relatively modal voice such as interest (`int`) and irritation (`irr`) perform best. But also hot anger (`col`) does not perform badly, though there is lots of shouting going on in these samples. Pleasure (`pla`) on the other hand, which often comes with a very breathy voice and some emotive bursts performs considerably worse. The training data was not especially balanced in respect to the distribution of emotions, though. A more detailed evaluation and interpretation of the

respective behaviour of the different emotions is thus left open for future investigations.

When having a look at absolute error values again it becomes clear that the resulting segmentations should not be used as is but need to be manually corrected. The amount of effort required for this mainly depends on the percentage of samples with an error of less than 20ms, as these are the ones that most probably would not need to be changed by a human labeler.

Our experience indicated that finally 2 minutes per sentence are a realistic estimate for the time span required for performing manual editing of the alignment results. This also includes taking notes whenever peculiarities in the voice quality (e.g. breathyness, flustering, creakyness) or segmental abnormalities such as affective bursts are observed. Cf. Appendix B and C.

Based on the current experience with annotating the corpus we do not consider that restricting the manual correction to the level of syllable boundaries in order to save time and effort is to be recommended. Though syllable level segmentations probably provide sufficient temporal resolution for a number of research tasks, e.g. for analysis of synchronisation on gestures and speech, we do not think that the time savings are worth the loss of information when dropping phone-level segmentation.

5 Conclusions

In this paper the work on the automatic phonetic segmentation of emotional speech from the GEMEP corpus was presented. On the one hand the highly uniform segmental content of GEMEP should facilitate this task. On the other hand it was an open question on how well the standard techniques – applying Hidden Markov Models trained on Mel Frequency Cepstral Coefficients – would deal with the various types of *emotional* speech in the corpus.

Both the subjective experience and the objective measures provided in this report show, that it was possible to come up with an aligner that produced satisfying results for about 80% of the data rather quickly, i.e., the amount of training data and thus the effort required for building a running alignment system could be kept within reasonable bounds. On the other hand it turned out that neither considerably huger amounts of training data nor the application of different training methods brought a further steep increase in quality. This means that most of the segmentations are handled nicely even with rather limited effort but a considerable amount of problematic cases remain resistant to further attempts of improvement.

Manual correction of the automatic segmentations remains a necessity for this kind of data. Performing this correction task has been an integral part of the work reported here. As a result manually validated segmentations for a significant portion of the GEMEP corpus have been made available for use in future studies. [Pirker 2007] is a first example of the kind of investigations made possible with the availability of phonetically segmented data.[5]

[5]Additional resources connected with this report, e.g., the README-files from the Appendices, can be accessed d at `http://www.ofai.at/~hannes.pirker/gemep`

6 Acknowledgements

I am very much indebted to Klaus Scherer and his group in Geneva for designing, creating and sharing their GEMEP corpus and especially to Tanja Bänziger for long standing and ongoing interaction and support. This work has been funded by the EU Network of Excellence HUMAINE (IST 507422) and by the Austrian Funds for Research and Technology Promotion for Industry (FFF 808818/2970 KA/SA). Financial support for OFAI is provided by the Austrian Federal Ministry of Science and Research and by the Federal Ministry of Transport, Innovation and Technology. This publication reflects only the author's views. The European Union is not liable for any use that may be made of the information contained herein.

References

[Bänziger et al. 2006] Bänziger T., Pirker H., Scherer K.: GEMEP - GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions, in Devillers L. et al. (eds.), Proceedings of LREC'06 Workshop on Corpora for Research on Emotion and Affect, May 23, Genoa, Italy, pp.15-19, 2006.

[Bänziger et al. 2007] Bänziger T., Scherer K.R.: Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus, in Paiva A. et al. (eds.), Affective Computing and Intelligent Interaction, Springer-Verlag Berlin Heidelberg, pp.476-484, 2007.

[Kohler 1994] Kohler K.: Lexica of the Kiel PHONDAT Corpus, Read Speech, Institut für Phonetik und Digitale Sprachverarbeitung, Universität Kiel, Vol.I, Arbeitsberichte Nr.27, 1994.

[Young et al. 2006] Young S., Evermann G., Kershaw D., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland P.: The HTK Book (version 3.4), Cambridge University Engineering Department, Cambridge UK, 2006.

[Pirker 2007] Pirker H.: Mixed Feelings About Using Phoneme-Level Models in Emotion Recognition, in Paiva A. et al. (eds.), Affective Computing and Intelligent Interaction, Springer-Verlag Berlin Heidelberg, pp.772-773, 2007.

[R Core Team 2007] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2007. [http://www.R-project.org]

A Appendix: README_transcripts

```
#########################################################################
## LABELLING SCHEME for the PHONETIC TRANSCRIPTION of the GEMEP CORPUS
##
## EXPLANATIONS and COMMENTS
#########################################################################
## Hannes Pirker OFAI, v 0.2 November 2007
#########################################################################


1.CANONICAL TRANSCRIPTION
-------------------------

In the labelling scheme the following canonical transcription for the
two sentence-types is used:

The canonical transcription for the two sentences is:

TYPE1: / _ n e k a l i b a m (@) _ s u d (@) m o l (e|E) n (@|A) _  /
TYPE2: / _ k u n s e m i n a _ l o d (@) _ b e l a m (@|A) _  /

These labels are compliant with SAMPA.

Special note on plosives: There are two tricky cases in {kun se
mina...}. First the actual start of the /k/ usually cannot be properly
determined as its occlusion-pause merges with the sentence initial
pause. Second, in {...lod belam...} /d/ and /b/ often merge into each
other. In this case a position more or less in the middle of the
resulting joined pause was chosen as 'boundary' between the two
sounds.

Special Notice on /e/ vs. /E/ i.e. tense vs. lax vowel: It turned out
that speakers articulate the /e/ in the final syllable in {molén} as
/E/. In the first phase of manual annotation this fact was ignored in
the labelling scheme and all appearances were uniformly annotated as
/e/. Only in later cycles of the annotation preocedure /e/ and /E/
were properly distinguished. Therefore tokens of /e/ in this syllable
can actually refer to /E/ in some samples, while /E/ should reflect
the degree of openness correctly.

Special Notice on SCHWA sounds: i) Due to the French origin of the
speakers the insertion of schwa-sounds (SAMPA: /@/) in syllable-final
position appears so frequently in the corpus, that these sounds are
treated as 'canonical', i.e. their appearance is not treated as
'insertion' of an non-canonical label.
```

ii) In a considerable number of cases (usually in sentence final position of {Ne kal ibam...}) the /@/ is so prominently articulated that it was decided to use the label /A/ (slightly against its use in SAMPA) for explicitly marking these occurrences as 'strong'. (These /A/ are very unevenly distributed among different speakers and emotions, BTW).

2. INSERTIONS and SUBSTITUTIONS

Deviations from the canonical transcription, are made explicit using the labelling scheme developed for the Phondat-Korpus at Kiel University (Kohler 1994).

"=-A" : insertion of an (non-canonical) segment A

"A-B" : replacement of (canonical) A by B

"A-:" : deletion of canonical A

Note: Due to technical reasons deletions are right now NOT explicitly marked in the transcription but if needed this information could be automatically inferred from the existing labels.

Most common cases:

/b-w/ : substitution of /b/ with /w/ : due to sloppy articulation.
{Ne kal ibam} becomes {Ne kal i*w*am}
articulation.

/n-d/ : substitution of /n/ with /d/ :
{Kun se ...} becomes {Ku*d*se ...}

/=-h/ : a /=-h/ is used as a marker for different sorts of interruption in the vocal flow which appears quite often. (A different encoding, namely /h_X_LAUGH/ is used for marking the appearance of an vocal occlusion in laughter, see below)

3. PECULIARITIES in SPEECH-SOUNDS

"_X" can be appended to any label in order to point out 'peculiarities' of this sound. This could e.g. be uncommon modes of pronunciation.

In many cases the nature of these 'peculiarity' is made more explicit
by appending another label to the "_X".

Special labels are:

"_X_H" :
  Aspiration. The marker "_H" is used for depicting both strong
  aspiration and high degree of breathyness. Most commonly found
  in '@_X_H' and'A_X_H', i.e. strongly aspirated sentence final
  'schwa' sounds. Note: Especially for /A_X_H/ the distinction
  from an emotive burst (e.g. a sigh) is not always clear-cut.

"_X_LAUGH" :
        Laughing. The sound is articulated in a 'laughing'
        manner. Most common token is 'h_X_LAUGH'. Another label used
        for marking laughter is 'BURST_LAUGH', see below.

"_X_h" :
        Interruption by /h/. In many of the samples a vowel that is
        interrupted by /h/ (i.e. either a '=-h' or a 'h_X_LAUGH') is
        also explicitly marked. I.e. an /e/ pronounced in an 'laughing
        mode' can be transcribed as /... e_X_h h_X_LAUGH e_X_h .../

"_X_BLAST" :
  The amplitude of the recording is blasting, i.e. the sound is
  clipped.

"_X_NOISE" :
Some audible intermingled noise.

"n_X_CUT" :
parts of initial /n/ in {ne kal ...} has been CUT i.e. its
START was cut off

"n_X_TRUNC":
parts of final /n/ in {... molen} has been TRUNCATED i.e. its
END was cut off

Voice quality:

In a VERY low number of cases out-standing peculiarities in the voice
quality or speaking mode were explicitly marked by appending the
following markers:

```
 "_X_FLUSTER"
 "_X_LARYNGALIZED"
 "_X_CREAKY"
 "_X_FALSETTO"


NON-SPEECH SOUNDS / NOISE / etc.
--------------------------------

"GARB" : 'GARBAGE': these are usually samples with SEGMENTAL content
that cannot be sensefully processed: typically at the start or the end
of the sentence. Most often these portions could be removed from the
sample without loss of information.

"FILLER" :
typically an /a/ or /@/ sound for marking a 'filled pause' or
a hesitation.

"NOISE" :
 non-speech sounds of insecure origin or typical disturbing
 sounds like clapping etc.

"BREATH" :
"Normal" breathing sound, clearly audible

---

"BURST" :
this subsumes all the kinds of non-speech sounds that were
perceived of bearing 'affective meaning'. 4 different classes
are distinguished: LAUGH, BREATH, NASAL, and X (others).

"BURST_LAUGH" : Laughing

"BURST_BREATH_(INHALE|EXHALE)" :
breathing sound produced by inhaling or exhaling and judged as
affective BURST.

"BURST_NASAL_ ..."  :
sound produced by nasal inhaling or exhaling

    BURST_NASAL_H : a moderate air OUTtake, i.e. a sort of nasal
                        aspiration on the sentence-final /@/.

    BURST_NASAL_SNEEZE : loud nasal air INTAKE or OUTtake
    BURST_NASAL_SNIFF  : a moderate nasal air INTAKE
```

"BURST_X" :
BURST that is none of class BREATH, NASAL, LAUGH.

The labelling convention for further specifying 'BURST_X' is to use
either uppercase letters only for naming a phenomenon or initial
uppercase followed by lowercase for providing a rough orthographic
impression of the sound produced.

Labels are:

BURST_X_SMACK :
sentence initial SMACK of the lips
BURST_X_COUGH :
coughing-like sound

BURST_X_Ah :
like sentence-final /A/ but with a strong aspiration!
Attention: Fuzzy boundaries to: @_X_H, A_X_H,
BURST_BREATH_EXHALE!

BURST_X_Ahhh :
acoustically reesembles BURST_X_Ah but is positionally clearly
distinguished as a sentence initial burst.

BURST_X_Hah :
acoustically and 'semantically' heterogeneous group.

BURST_X_Mmm :
/m/ (purring sound) typically found in samples of pleasure

BURST_X_Bhhh :
sentence-initial burst, strongly aspirated /b/

REFERENCES

Kohler K.: Lexica of the Kiel PHONDAT Corpus, Read Speech, Institut
fuer Phonetik und Digitale Sprachverarbeitung, Universitaet Kiel, Vol.I,
Arbeitsberichte Nr.27, 1994.

SAMPA: SAMPA computer readable phonetic alphabet,
[http://www.phon.ucl.ac.uk/home/sampa/index.html, last visited: 10.10.2007]

B Appendix: README_remarks

```
01mep12G| mixed text! "nekalibam sud bela!"
01pla111| sud /m-v/olen
02irr112| final /@/ very long and breathy: marked as /A_X_H/
02irr212| very breathy/noisy
02irr313| [ibOm] ! marked as /i b a-O m/
02irr313| TRUNCATED: initial syllable: /ne/ is missing: instead: BREATH _ kal ...
02peu212| /o/ (almost?) missing
03irr112| VERY low volume ... creaky and breathy
03mep116| EXTREMELY low volume, thus NOISY in the amplified version
03mep119| EXTREMELY low (+91, -58) + thus NOISY in the amplified version
04int112| NOISE recording errors (2 scratchy sounds) - alignment almost o.k.
04int113| TOTALLY CORRUPTED: only consists of a burst ==> completely removed!
04joi415| sloppy articulation
04joi416| TRUNCATED first syllable shouted, rest of sent low and incomplete
04tri311| /s u d m _ m o / : a hesitative pause IN /m/
04tri312| exremely low volume in sudmole(n): almost not audible
05amu111| CORRUPTED. other speaker audible in background
05amu112| TOO much laughter & over-acting to be useful?
05irr312| /m/ is exremely "pressed": marked as /m_X/
05joi213| LAUGHTER
05sou311| BURST_X_Ahhh ne kal ibam...
06amu111| LAUGHTER
06amu112| LAUGHTER
06att112| TRUNCATED: starts with /ka li bam/
06col215| TRUNCATED: initial n is missing and also /e/ is truncated
06int411| untypical long pause: sudm _PAUSE_ molen
06int416| rising intonation contour (question rise)
06joi112| final /@/ is almost an emotive burst. transcribed as /A_X/
07pla115| ends in a nasal burst? transcribed as /A_X_H/
07pla212| /e/ in /ne/ is actually more a high pitched /n/?
08amu412| n_X_TRUNC
08peu128| stuttering: /ku k u k u d /
08pla315| LAUGHTER very prominent burst + laughter: "Ooho haha ne kal ..."
09amu112| LAUGHTER
09hon115| TRUNCATED: initial /ne/ is missing
09int112| very low volume. transcribed as /s u m o l/
09irr212| untypical syllabification: / nekal _ ibam /
09joi411| TRUNCATED: missing: initial syll, final /n/, pronounciation: /k a-e/
09sou124| DELETION missing /m/
09tri418| /n e k-g a l i b a m /
10amu112| LAUGHTER
10deg111| /i/ not audible. Marked /l_X b-w/. sloppy ('drunk-like') articulation
10fie112| initial SMACK of the lips
```

C Appendix: README_voice_quality

```
# ======================
# README_voice_quality
# ======================

## Hannes Pirker OFAI, 2007 This file contains remarks on
## peculiarities in the sound-files.  The listing under the different
## categories is NOT a complete enumeration, i.e. view it a collection
## of examples for some 'non-standard' modes of articulation.

## SEE ALSO: README_remarks

## - This file can also be transformed to a TABLE:
## 'README_voice_quality.tab' with the following command:

## cat README_voice_quality | egrep -v '^#' |
## perl -e 'LINE:while(<>) {chomp; if(s/^INFO_//) {$info=$_}; if
## (m/^\d\d/) {($f,$comment)=split(/\|/);
## print("$f|$info|$comment\n")}}' | perl -pe 's/ \|/\|/g;' | sort -n
## -t\| -k1 >! README_voice_quality.tab

# Sample output:
#01amu314|HIGH_PITCH_SHOUTING| a_X_BLAST
#01att126|VOLUME_LOW|
#01col311|HIGH_PITCH_SHOUTING|
#01col313|HIGH_PITCH_SHOUTING|
## ------------------------------------

# =======
INFO_VOLUME_HIGH
# ========

03peu112
09joi311
09joi312
03joi111| extremely long l /E/ n (2.4 sec!!!!)
03joi112| extremely long l /E/ n (1.7 sec!!!!)

# ==========
INFO_HIGH_PITCH_SHOUTING
# ==========

01amu314| a_X_BLAST
01col311
01col313
01joi114
```

```
02col124
03col311
03col312
03des314
03des315
04joi211
04joi311| e_X_BLAST
05col12A
05peu311
05peu314
07peu112
07peu117
07peu126
09col112
09col113
09col124
09joi313| only /e/ in /len/ VERY high pitched, high energy, extremely long
09peu311

# ========
INFO_VOLUME_LOW
# =========

01mep114
01att126
03mep111
03mep112| low f0
03mep113| low f0
04pla124
04tri124
04tri125
07pla125
07tri126
08tri125
09int125
09pla125
10hon111
10hon112| beating-noise in the /o/ sound

# ==========
INFO_LOW_PITCH_BREATHY_VOICE
# ==========

01sou123
02adm126
04adm124
04adm125
```

```
04hon1110| very breathy in the end
03pla123
03sur111
03sur112
03tri123
05pla121
06sur124
08fie116
08tri1114
08tri111
08tri127
10hon113| end of sentence (sud molen) is flustered
10sou211

# =====
INFO_BREATHY_VOICE
# =====

10sou123| very breathy+noisy, high effort

# ==========
INFO_HIGH_PITCH
# ==========

04amu1214
06des121
06des123
06peu124| extrem high - over 800Hz !
06joi125| falsetto? at least very HIGH pitch + loud
07amu313
07peu125
08joi118| LAUGHTER
08joi119
08joi121
08pla128| SINGING-like (long regions with steady pitch)
08pla211| only final /lam/ gets very high
08peu112
09amu123
10joi313| m o _ E (nTrunc) /E/ extreme high+falsetto !
10amu121

# ==========
INFO_SOFT_VOICE
# ==========

01mep115
01sur113
```

```
02adm111
02deg111
02hon111
02hon113
02hon1110
02hon119
02joi414
02pla125
07att128
08mep113
08mep116
08mep123
08pla115
09adm125
09hon112
09hon113
10adm129
10tri121
04adm116

# ==========
INFO_FLUSTERING_VOICE
# ==========
01des212| last syllable almost inaudible: /len/
02joi412
02sou412
03inq214| strong aspiration + breathing
03inq216| strong aspiration + breathing
03mep116
03mep119
04pla114
05hon114
06sur112| TRUNCATED: initial /ne/ missing; /kal ibam/ flustered
06sur114
06sur115
06sur116| final /e n/ barel visible, but still audible
08des117| parts are flustered
08att123
08int3112
08mep124
08sou113
08sou113| initial /n/ is CREAKY and rather long (280ms)
08sou114
08sur117
08pla119
09sou212
10adm1111| strong aspiration in parts of file: marked as /n_X_H e_X_H/ etc.
```

```
10adm115
10hon1116
10hon1119
10hon1119| start of sentence: FLUSTERING

# ==========
INFO_HIGH_EFFORT_VOICE
# ==========
04deg112
05peu124
09deg111
09peu315| extremely long /E/ in l/E/n (1.6sec), high f0 of ca. 600Hz in /E/

# ==========
INFO_CREAKY VOICE
# ==========
02deg125
03des216
03int121
10deg113| creaky at start: /ne kal/

# ==========
INFO_LARYNGALIZED VOICE
# ==========
02deg112
02deg123
05deg114
03irr117
08inq218| laryngalization rather visible (spectrum) than audible
```

D Appendix: Description of Training- and Test-Corpora

```
## ------------------------------------------------------
## README_corpora
## Overview of the different subsets of the GEMEP-corpus
## used for TRAINING and TESTING phonetic segmentation
## ------------------------------------------------------

-------------
I) _ALL
-------------

The whole set of 3818 recordings of type1 and type2 sentences available
in GEMEP.

============ START statistics of _ALL =============
== Produced by: actor_statistics.csh
== Directory:   /hp/GENEVE_DATA/SFS/_ALL
== Date:        2007-11-23
=============

TOTAL number of files:  3818

-------------------------
1. ABSOLUTE numbers
-------------------------

------ TYPE -------------
   1    2
2741 1077
------ ACTOR ------------
 01  02  03  04  05  06  07  08  09  10
301 304 326 368 333 309 450 545 373 509
------ EMO --------------
adm amu att col deg des fie hon inq int irr
 83 179  89 210  83 242 316 126 378 319 271
joi mep peu pla sou sur tri
215  92 250 263 346  96 260

------ REGULATION TYPE---
   1    2    3    4
2334  376  459  649
------ REPETITIONS ------
  1   2   3   4   5   6   7   8   9  10  11
478 478 462 414 338 277 225 183 141 105  76

11  12  13  14  15  16  17  18  19  20
```

```
76  59  44  29  25  20  17  13   9   6

 21  22  23   A   B   C   D   E   F   G
  4   1   1 180  82  46  34  23  18  14

H   I   J   K   L   M   N   O   P
6   3   1   1   1   1   1   1   1

mean number of REPETITIONS:  7.954167

------------------------
 2. RELATIVE numbers in % (rounded)
------------------------

------ TYPE % ----------
 1  2
72 28
---- -ACTOR % ----------
01 02 03 04 05 06 07 08 09 10
 8  8  9 10  9  8 12 14 10 13
------ EMO % ------------
adm amu att col deg des fie hon inq int irr
  2   5   2   6   2   6   8   3  10   8   7
joi mep peu pla sou sur tri
  6   2   7   7   9   3   7

------ REGULATION TYPE %
 1  2  3  4
61 10 12 17
------ REPETITIONS % ----
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17
13 13 12 11  9  7  6  5  4  3  2  2  1  1  1  1  0
18 19 20 21 22 23  A  B  C
 0  0  0  0  0  0  5  2  1
 D  E  F  G  H  I  J  K  L  M  N  O  P
 1  1  0  0  0  0  0  0  0  0  0  0  0
------------ END statistics of _ALL ---------------


-------------------------
II) TRAIN_n1313
-------------------------

*TRAIN_n1313* : ALL files for which a manual segmentation is
available. (status of November 2007).

This is further segmented into a TEST- and several TRAINING-corpora
```

```
============ START statistics of TRAIN_n1313 =============
== Produced by: actor_statistics.csh
== Directory:   /hannes/GENEVE/MANUAL_LAB_tutti/TRAIN_n1313
== Date:        2007-11-27
=============

TOTAL number of files:  1313

-------------------------
 1. ABSOLUTE numbers
-------------------------

------ TYPE -------------
  1   2
973 342
------ ACTOR ------------
 01  02  03  04  05  06  07  08  09  10
134 124 122 124 155 157 122 122 125 130
------ EMO --------------
adm amu att col deg des fie hon inq int irr
 34  81  38  81  34  85  83  34 112 143 104
joi mep peu pla sou sur tri
 83  36  83  82  83  36  83
------ REGULATION TYPE---
  1   2   3   4
865 146 151 153
------ REPETITIONS ------
  1   2   3   4   5   6   7   8   9  10  11  12  13  14
225 270 241 180 119  78  61  35  28  19  12  17   4   3

 15  16  17  18  19  20
  3   5   2   2   2   1

 21   A   B   E   F   G
  1   2   1   2   1   1

mean number of REPETITIONS:  2.745303

------------------------
 2. RELATIVE numbers in % (rounded)
------------------------

------ TYPE % -----------
 1  2
74 26
----- ACTOR % -----------
```

```
01 02 03 04 05 06 07 08 09 10
10  9  9  9 12 12  9  9 10 10
------ EMO % ------------
adm amu att col deg des fie hon inq int irr
  3   6   3   6   3   6   6   3   9  11   8
joi mep peu pla sou sur tri
  6   3   6   6   6   3   6

------ REGULATION TYPE %
 1  2  3  4
66 11 11 12
------ REPETITIONS % ----
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17
17 21 18 14  9  6  5  3  2  1  1  1  0  0  0  0  0
18 19 20 21  A  B  E  F  G
 0  0  0  0  0  0  0  0  0

================ END statistics of TRAIN_n1313 ====================

-----------------------
III) TRAIN_892_of_n1313
-----------------------

Contains all samples from TRAIN_n1313 MINUS the samples from the TEST-set,
i.e. all samples MINUS repetitions 3+4
-----------------


============ START statistics of TRAIN_n892_of_n1313 =============
== Produced by: actor_statistics.csh
== Directory:   hannes/GENEVE/MANUAL_LAB_tutti/TRAIN_n1313/TRAIN_n892_of_n1313
== Date:        2007-11-23
=============

TOTAL number of files:  894


-------------------------
 1. ABSOLUTE numbers
-------------------------

------ TYPE -------------
  1   2
685 209
------ ACTOR ------------
 01  02  03  04  05  06  07  08  09  10
 80  79  84  77 117 110  82  87  81  97
```

```
------ EMO --------------
adm amu att col deg des fie hon inq int irr joi mep peu pla sou sur tri
 27  54  24  56  21  52  54  23  83  93  75  48  23  57  57  61  25  61
------ REGULATION TYPE---
  1   2   3   4
587  87 103 117
------ REPETITIONS ------
  1   2   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21   A
225 270 119  78  61  35  28  19  12  17   4   3   3   5   2   2   2   1   1   2
  B   E   F   G
  1   2   1   1
mean number of REPETITIONS:  1.935065



------------------------
 2. RELATIVE numbers in % (rounded)
------------------------

------ TYPE % ----------
 1  2
77 23
---- -ACTOR % ----------
01 02 03 04 05 06 07 08 09 10
 9  9  9  9 13 12  9 10  9 11
------ EMO % ------------
adm amu att col deg des fie hon inq int irr joi mep peu pla sou sur tri
  3   6   3   6   2   6   6   3   9  10   8   5   3   6   6   7   3   7
------ REGULATION TYPE %
 1  2  3  4
66 10 12 13
------ REPETITIONS % ----
 1  2  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21  A  B  E  F  G
25 30 13  9  7  4  3  2  1  2  0  0  0  1  0  0  0  0  0  0  0  0  0  0

------------ END statistics of TRAIN_n892_of_n1313 ---------------

------------------------
IV) TEST_n421_of_n1313
------------------------

In order to get a simple and consistent way of identifying files to be
used for testing, it was decided to always reserve samples of repetion
number #3 and #4 (i.e. the third and fourth repetition) for training.

----------------------------------------
Contains ALL and ONLY reptitions  3 + 4 from TRAIN_n1313
```

```
----------------------------------------

============ START statistics of TEST_n421_of_n1313 =============
== Produced by: actor_statistics.csh
== Directory:   /hannes/GENEVE/MANUAL_LAB_tutti/TRAIN_n1313/TEST_n421_of_n1313
== Date:        2007-11-23
=============

TOTAL number of files:  421

-------------------------
 1. ABSOLUTE numbers
-------------------------

------ TYPE -------------
  1   2
288 133
------ ACTOR ------------
01 02 03 04 05 06 07 08 09 10
54 45 38 47 38 47 40 35 44 33
------ EMO --------------
adm amu att col deg des fie hon inq int irr
  7  27  14  25  13  33  29  11  29  50  29
joi mep peu pla sou sur tri
 35  13  26  25  22  11  22
------ REGULATION TYPE---
  1   2   3   4
278  59  48  36
------ REPETITIONS ------
  3   4
241 180
mean number of REPETITIONS:  1.358065

------------------------
 2. RELATIVE numbers in % (rounded)
------------------------

------ TYPE % ----------
 1  2
68 32
---- -ACTOR % ----------
01 02 03 04 05 06 07 08 09 10
13 11  9 11  9 11 10  8 10  8
------ EMO % ------------
adm amu att col deg des fie hon inq int irr
  2   6   3   6   3   8   7   3   7  12   7
 joi mep peu pla sou sur tri
```

```
   8   3   6   6   5   3   5
------ REGULATION TYPE %
 1  2  3  4
66 14 11  9
------ REPETITIONS % ----
 3  4
57 43
------------ END statistics of TEST_n421_of_n1313 ---------------


-----------------------
V) TRAIN_n892_of_n1313
-----------------------
-----------------
Contains all samples from TRAIN_n1313 MINUS the samples from the TEST-set,
i.e. all samples MINUS repetitions #3 and #4
-----------------

============ START statistics of TRAIN_n892_of_n1313 =============
== Produced by: actor_statistics.csh
== Directory:   /hannes/GENEVE/MANUAL_LAB_tutti/TRAIN_n1313/TRAIN_n892_of_n1313
== Date:        2007-11-23
=============

TOTAL number of files:  894

-------------------------
 1. ABSOLUTE numbers
-------------------------

------ TYPE -------------
  1   2
685 209
------ ACTOR ------------
 01  02  03  04  05  06  07  08  09  10
 80  79  84  77 117 110  82  87  81  97
------ EMO --------------
adm amu att col deg des fie hon inq int irr joi mep peu pla sou sur tri
 27  54  24  56  21  52  54  23  83  93  75  48  23  57  57  61  25  61
------ REGULATION TYPE---
  1   2   3   4
587  87 103 117
------ REPETITIONS ------
  1   2   5   6   7   8   9  10  11  12  13  14  15
225 270 119  78  61  35  28  19  12  17   4   3   3

16  17  18  19  20  21   A
```

```
 5   2   2   2   1   1   2

  B   E   F   G
  1   2   1   1
mean number of REPETITIONS:  1.935065

------------------------
 2. RELATIVE numbers in % (rounded)
------------------------

------ TYPE % ----------
 1  2
77 23
---- -ACTOR % ----------
01 02 03 04 05 06 07 08 09 10
 9  9  9  9 13 12  9 10  9 11
------ EMO % ------------
adm amu att col deg des fie hon inq int irr
  3   6   3   6   2   6   6   3   9  10   8

joi mep peu pla sou sur tri
  5   3   6   6   7   3   7

------ REGULATION TYPE %
 1  2  3  4
66 10 12 13
------ REPETITIONS % ----
 1  2  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21  A  B  E  F  G
25 30 13  9  7  4  3  2  1  2  0  0  0  1  0  0  0  0  0  0  0  0  0  0
------------ END statistics of TRAIN_n892_of_n1313 ---------------



----------------------
VI) TRAIN_n503_of_n1313
----------------------

This set was created from TRAIN_n892_of_n1313 by removing all
repetions #2 and #5

cd TRAIN_n892_of_n1313;
mkdir ~/TRAIN_nXXX_of_n1313
cp *.bilab *.lab *.sfs ~/TRAIN_nXXX_of_n1313 ; pushd ~/TRAIN_nXXX_of_n1313
\rm  ???????2.*; \rm  ???????5.*

============ START statistics of TRAIN_n503_of_n1313 =============
== Produced by: actor_statistics.csh
```

```
== Directory:   /hannes/GENEVE/MANUAL_LAB_tutti/TRAIN_n1313/TRAIN_n503_of_n1313
== Date:        2007-11-23
=============

TOTAL number of files:  503

-------------------------
 1. ABSOLUTE numbers
-------------------------

------ TYPE -------------
  1   2
395 108
------ ACTOR ------------
01 02 03 04 05 06 07 08 09 10
36 34 44 44 70 57 48 64 43 63
------ EMO --------------
adm amu att col deg des fie hon inq int irr
 14  32  10  32  11  29  29  15  50  57  40

joi mep peu pla sou sur tri
 24  13  27  31  37  13  39

------ REGULATION TYPE---
  1   2   3   4
330  39  58  76
------ REPETITIONS ------
  1   6   7   8   9  10  11  12  13  14  15  16
223  78  61  35  28  19  12  17   4   3   3   5

17  18  19  20  21   A   B   E
 2   2   2   1   1   2   1   2

  F   G
  1   1
mean number of REPETITIONS:  1.466472

------------------------
 2. RELATIVE numbers in % (rounded)
------------------------

------ TYPE % ----------
 1  2
79 21
---- -ACTOR % ----------
01 02 03 04 05 06 07 08 09 10
 7  7  9  9 14 11 10 13  9 13
```

```
------ EMO % ------------
adm amu att col deg des fie hon inq int irr
  3   6   2   6   2   6   6   3  10  11   8

joi mep peu pla sou sur tri
  5   3   5   6   7   3   8

------ REGULATION TYPE %
 1  2  3  4
66  8 12 15
------ REPETITIONS % ----
 1  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21  A  B  E  F  G
44 16 12  7  6  4  2  3  1  1  1  1  0  0  0  0  0  0  0  0  0  0
------------ END statistics of TRAIN_n503_of_n1313 ---------------

------------------------
VII) TRAIN_n225_of_n1313
------------------------

Another selection of manual 1313 for training.
Reason: show how results change with training-size

This set was created from TRAIN_n892_of_n1313 by only selecting
repetition #1

cd TRAIN_n892_of_n1313;
mkdir ~/TRAIN_xxx_of_n1313
cp ???????1.*  ~/TRAIN_xxx_of_n1313 ; pushd ~/TRAIN_xxx_of_n1313

============ START statistics of TRAIN_n225_of_n1313 =============
== Produced by: actor_statistics.csh
== Directory:   /hannes/GENEVE/MANUAL_LAB_tutti/TRAIN_n1313/TRAIN_n225_of_n1313
== Date:        2007-11-23
=============

TOTAL number of files:  225

-------------------------
 1. ABSOLUTE numbers
-------------------------

------ TYPE -------------
  1   2
177  48
------ ACTOR ------------
01 02 03 04 05 06 07 08 09 10
23 18 28 23 35 27 15 16 22 18
```

```
------ EMO --------------
adm amu att col deg des fie hon inq int irr
  6  23   5  18   6  12  13   3  16  21  15
joi mep peu pla sou sur tri
 16   4  15  18  13   5  16
------ REGULATION TYPE---
  1   2   3   4
150  24  26  25
------ REPETITIONS ------
  1
225
mean number of REPETITIONS:  1

------------------------
 2. RELATIVE numbers in % (rounded)
------------------------

------ TYPE % ----------
 1  2
79 21
---- -ACTOR % ----------
01 02 03 04 05 06 07 08 09 10
10  8 12 10 16 12  7  7 10  8
------ EMO % ------------
adm amu att col deg des fie hon inq int irr joi mep peu pla sou sur tri
  3  10   2   8   3   5   6   1   7   9   7   7   2   7   8   6   2   7
------ REGULATION TYPE %
 1  2  3  4
67 11 12 11
------ REPETITIONS % ----
  1
100
------------ END statistics of TRAIN_n225_of_n1313 ---------------

-------------------------
VIII) TRAIN_n119_of_n1313
-------------------------

Another selection of manual 1313 for training.
Reason: show how results change with training-size
Just take all repetitions #5

cp TRAIN_n892_of_n1313/???????5.* TRAIN_XXXX/
mv TRAIN_XXXX TRAIN_n119

============ START statistics of TRAIN_n119 =============
== Produced by: actor_statistics.csh
```

```
== Directory:   /hannes/GENEVE/MANUAL_LAB_tutti/TRAIN_n1313/TRAIN_n119
== Date:        2007-11-23
=============

TOTAL number of files:  119

-------------------------
 1. ABSOLUTE numbers
-------------------------

------ TYPE -------------
 1  2
87 32
------ ACTOR ------------
01 02 03 04 05 06 07 08 09 10
11  9 16 11 11 15 13  8 11 14
------ EMO --------------
adm amu att col deg des fie hon inq int irr
  6   1   6   6   4   8   7   3   9  13  10

joi mep peu pla sou sur tri
  4   3  11  11   7   5   5

------ REGULATION TYPE---
 1  2  3  4
79 15 11 14
------ REPETITIONS ------
  5
119
mean number of REPETITIONS:  1

------------------------
 2. RELATIVE numbers in % (rounded)
------------------------

------ TYPE % ----------
 1  2
73 27
---- -ACTOR % ----------
01 02 03 04 05 06 07 08 09 10
 9  8 13  9  9 13 11  7  9 12
------ EMO % ------------
adm amu att col deg des fie hon inq int irr joi mep peu pla sou sur tri
  5   1   5   5   3   7   6   3   8  11   8   3   3   9   9   6   4   4
------ REGULATION TYPE %
 1  2  3  4
66 13  9 12
```

```
------ REPETITIONS % ----
  5
100
------------ END statistics of TRAIN_n119 ---------------

-------------------------
IX) TRAIN_n61_of_n1313
-------------------------

Make another selection of manual 1313 for training.
Just take all repet=#7

cp TRAIN_n892_of_n1313/????????.* TRAIN_n61_of_n1313/

============ START statistics of TRAIN_n61_of_n1313 =============
== Produced by: actor_statistics.csh
== Directory:   /hannes/GENEVE/MANUAL_LAB_tutti/TRAIN_n1313/TRAIN_n61_of_n1313
== Date:        2007-11-23
=============

TOTAL number of files:  61

-------------------------
 1. ABSOLUTE numbers
-------------------------

------ TYPE -------------
 1  2
46 15
------ ACTOR ------------
01 02 03 04 05 06 07 08 09 10
 3  4  2  8  5  4  6 11  4 14
------ EMO --------------
adm amu att col deg des fie hon inq int irr
  2   3   1   2   2   7   3   2   5   6   8

joi peu pla sou sur tri
  2   4   2   5   2   5

------ REGULATION TYPE---
 1  2  3  4
43  2  7  9
------ REPETITIONS ------
 7
61
mean number of REPETITIONS:  1
```

```
------------------------
 2. RELATIVE numbers in % (rounded)
------------------------

------ TYPE % ----------
 1  2
75 25
---- -ACTOR % ----------
01 02 03 04 05 06 07 08 09 10
 5  7  3 13  8  7 10 18  7 23
------ EMO % ------------
adm amu att col deg des fie hon inq int irr joi peu pla sou sur tri
  3   5   2   3   3  11   5   3   8  10  13   3   7   3   8   3   8
------ REGULATION TYPE %
 1  2  3  4
70  3 11 15
------ REPETITIONS % ----
  7
100
------------ END statistics of TRAIN_n61_of_n1313 ---------------
```

E Appendix: Alignment grammar

```
/* in order to make this an exekutable grammar file use:
   HParse gemep_typ1_and_t2.grm gemep_typ1_and_t2.net

   foreach i (*.grm)
     echo "$i ..."; HParse $i $i:r.net
   end

 T1: Ne kal ibam sud molen!
 T2: Kun se mina lod belam?
*/

(
 ([T1.PAUSSTART] T1.N1 T1.E1 [T1.SIL1] T1.K T1.A1 T1.L1 [T1.SIL2]
                 T1.I T1.B T1.A2 T1.M1 [T1.SCHWA_m] [T1.SIL3]
                 T1.S T1.U [T1.D] [T1.SCHWA_d] [T1.SIL4]
                 T1.M2 T1.O T1.L2 T1.E2 [T1.N2 [T1.SCHWA_n]]
   T1.PAUSEND)
 |
([T2.PAUSSTART] T2.K T2.U (T2.N1|T2.D0) T2.S T2.E1 T2.M1 T2.I T2.N2 T2.A1
                T2.L1 T2.O T2.D T2.B T2.E2 T2.L2 T2.A2 [T2.M2 [T2.SCHWA_m]]
 T2.PAUSEND)
)
```

F Appendix: Source for Dictionary-Generation: *ne kal ibam...*

```
/* ============ from: nekalibam.source.dict  ======================================*/
/* perl program for adding TYP + EMO - labels to the 'naked'  file */

/* choose : TYP=(T1.|"")  and emo=(1|0) */
/* perl -e '$TYP="T1."; $emo = '1'; @emo=("adm","amu","att","col","deg","des","fie","

/* last edited hp 6MAR2007: perl script now adss "T1." 
/* last edited hp 28FEB2007: commented out old NE KAL ...  */
/* last edited hp 19FEB2007: minor adjustments */
/* last edited hp 23AUG2006: include unique SIL-names and FIL -> */
/*                intended for use with nekalibam.grm/nekalibam.net */

/* ([PAUSSTART] N1 E1 [SIL1] K A1 L1 [SIL2] I B A2 M1 [SCHWA_M] [SIL3] */
/*              S U D [SCHWA_D] [SIL4] M O L2 E2 N2 [SCHWA_N] PAUSEND) */

N1 n%e
N1 n_X_CUT%e

E1 e%k

K k%a

A1  a%l

L1 l%i
L1 l%_

I i%b

B b%a
B w%a
B b-w%a

A2 a%m

M1 m%s
M1 m%_
M1 m%@

S s%u
/* for bootstrap! */
S s%u_TRAINED_ON_n2940

U u%d
```

```
U u%m

D d%m
D d%_
D d%@

M2 m%o

O o%l

L2 l%e

E2 e%n
E2 E%n
/* sometimes final n is missing: */
E2 E%_
E2 e%_

N2 n%_
N2 n%@


/* map all silences to ONE model ? */
/* _ _ */
/* n_X_TRUNC n%_ */
/* _%n _ */
/* _%END _ */
PAUSSTART _%n
PAUSSTART _%B  B%_ _%n
/* AUG 2007: B renamed to BREATH */
PAUSSTART _%BREATH  BREATH%_ _%n
PAUSSTART _%NOISE NOISE%_ _%n

PAUSEND _%END
/* PAUSEND _%SOB SOB%_ _%END */

/* SIL */
/* for savety-reasons: add a _%n to each SIL */
SIL1 _%k
SIL1 _%n
/* as fallback! */
SIL2 _%i
SIL2 _%n
/* as fallback! */
SIL3 _%s
SIL3 _%n
/* as fallback! */
```

```
SIL4 _%m
SIL4 _%n
/* as fallback! */

/* frueher FILL */
FILL3 @%s
FILL3 @%_
FILL4 @%m
FILL4 @%_
FILLEND @%_
FILLEND A%_
/* jetzt SCHWA */
SCHWA_m @%s
SCHWA_m @%_
SCHWA_d @%m
SCHWA_d @%_
SCHWA_n @%_
SCHWA_n A%_

/* ========= end   of nekalibam.source.dict  ========== */
```

G Appendix: Source for Dictionary-Generation: *kun se mina...*

```
/* ============ from: kunsemina.source.dict ========================*/
/* perl program for adding TYP + EMO - labels to the 'naked'  file */

/* last edited hp 21AUG2007:  */
/* last edited hp 26FEB2007: derived from nekalibam.proto_per_emo.naked.dict */

/* ([T2.PAUSSTART] T2.K T2.U T2.N1 [T2.SIL_n] T2.S T2.E1 [T2.SIL_e]       */
/*     T2.M1 T2.I T2.N2 T2.A1 [T2.SIL_a] T2.L1 T2.O T2.D [T2.SCHWA_d]     */
/*     T2.[SIL_d] T2.B T2.E2 T2.L2 T2.A2 T2.M2 [T2.SCHWA_m] T2.PAUSEND) ) */

K k%u
K k%a

U u%n
U u%d
U u%m

/** D0 as alternative to N1**/
D0 d%s
/** D0 as alternative to N1**/
D0 n-d%s

/** D0 : because there are not enough training samples for EMO-specific D0 : */
/*       add mapping to d%m as a remedy ;-( **/
D0 d%m

N1 n%s
N1 n%_

S s%e

E1 e%m
E1 e%n
E1 e%_

M1 m%i

I i%_
I i%n

/* N */
N2 n%a

A1 a%l
```

```
A1 a%m

L1 l%o
L1 l%i
L1 l%e

O o%d
O o%l

D d%b
D d%_
D d%@
D d%m

B b%e
B b%a

/* E */
E2 e%l

/* L */
L2 l%a

/* A */
A2 a%m

M2 m%_
M2 m%@

/* --------------------------- */
/* map all silences to ONE model ? */
/* _ _ */
/* n_X_TRUNC n%_ */
/* _%n _ */
/* _%END _ */
PAUSSTART _%k
PAUSSTART _%B  B%_ _%k
/* AUG 2007: B renamed to BREATH */
PAUSSTART _%BREATH  BREATH%_ _%k
/* insufficient number of trainig-samples for NOISE ?? */
PAUSSTART _%NOISE NOISE%_ _%k
PAUSEND _%END
/* PAUSEND _%SOB SOB%_ _%END */

/* SIL */
/* for safety-reasons: add a _%n to each SIL */
SIL_n _%s
```

```
SIL_n _%END
SIL_n _%n /* as fallback! */

SIL_e _%m
SIL_e _%n /* as fallback! */

SIL_a _%l
SIL_a _%k
SIL_a _%m
SIL_a _%n /* as fallback! */

SIL_d _%b
SIL_d _%m
SIL_d _%n /* as fallback! */

/* frueher FILL  jetzt SCHWA */
SCHWA_m @%s
SCHWA_m @%_
SCHWA_d @%m
SCHWA_d @%_
/* ------------------------- */

/* ========= end   of kunsemina.source.dict  ========== */

/* =============== HACK! in order for typ2 kunsemina to work even when
training only was performed on typ1 nekal ibam : mapping from
nekalibam-labels to kunsemina-labels */

K k%a

U u%d
U u%m

N1 n%e
N1 n%_
N1 n%@

S s%u

E1 e%k
E1 e%n

M1 m%o
M1 m%_

I i%b
```

```
N2 n%e

A1 a%m
A1 a%l

L1 l%i
L1 l%e

O o%l

D d%m
D d%_
D d%@

B b%a

/* E */
E2 e%k
E2 e%n

/* L */
L2 l%i
L2 l%E
L2 l%e

/* A */
A2 a%m

M2 m%s
M2 m%_
M2 m%@

/* --------------------------- */

/* map all silences to ONE model ? */
PAUSSTART _%n
PAUSSTART _%k
PAUSSTART _%B  B%_ _%n
/* insufficient number of trainig-samples for NOISE ?? */
PAUSSTART _%NOISE NOISE%_ _%n

PAUSEND _%END
/* PAUSEND _%SOB SOB%_ _%END */

/* SIL */
/* for savety-reasons: add a _%n to each SIL */
SIL_n _%s
```

```
SIL_n _%END
/* SIL_n as fallback! */
SIL_n _%n

SIL_e _%k
/* SIL_n as fallback! */
SIL_e _%n

SIL_a _%k
SIL_a _%m
/* SIL_n as fallback! */
SIL_a _%n

SIL_d _%m
/* SIL_n as fallback! */
SIL_d _%n

SCHWA_m @%s
SCHWA_m @%_
SCHWA_d @%m
SCHWA_d @%_
/* ------------------------ */

/* ========= end   of kunsemina.proto.dict  ========== */
```