A German Viseme-Set for Automatic Transcription of Input Text Used for Audio-Visual-Speech-Synthesis

Christian Weiss, Bianca Aschenberner

Institute for Communication Research & Phonetics (IKP)
University of Bonn, Germany

{cwe,bas}@ikp.uni-bonn.de

10.21437/Interspeech.2005-777

Abstract

In this paper, we introduce a German viseme inventory for visemically transcribing text according to phonetic transcribtion. A viseme set like the one presented in this work is essential for speech-driven audio-visual synthesis due to the fact that the selection of appropriate video segments is based on the visemically transcribed input text.

For text-to-speech synthesis, a transcription of the input text into the phonemic representation is used, in order to avoid ambiguous meanings and to acquire the correct pronunciation of the underlying input text and to serve as labels in unit-selection-based synthesis systems. Likewise, the visual synthesis requires a transcription that represents - analogue to the phonemes - the visual counterpart which is called viseme in related literature and which also serves as a unit label in our data-driven video-realistic audio-visual synthesis system.

We worked out an inventory of German viseme classes in a SAMPA-like labelling and trained a model for automatic visemic transcription of given input text.

1. Introduction

Visemic transcripton is of text is used in audio-visual speech recognition for a longer period of time especially for automatic lip-reading. In that sense audio-visual synthesis becomes more and more interesting to both research and industry, because the fusion of the auditory and visual representation provides a human-machine interface being much more natural. For more than a decade, researchers have been exploring and experimenting with the relationship between speech and facial expressions that correspond to articulation, such as McGurk, McDonald [12] or Massaro [11]. They showed that besides the acoustic cues, the visual cues such as lip movements are also crucial to the speech perception and that the facial gesture plays a major role in face-to-face-communication as well as in Human-Machine-Interaction by improving the intelligibility of a spoken utterance. Consequently, this research is considered essential in numerous areas such as speech-reading education, E-commerce, customer relations as well as health services. To make the additional visual synthesis be a remarkable improvement of the existing speech synthesis systems, the quality of the visual information is imaginably important [3, 10]. Examples of computer animated "Talking Heads" which produce audio and lip synchronized speech can already be found in various applications [12], whereas the two major approaches currently are the model-based approach of building a Talking Head such as Baldi and Synface [17, 20] and the data-driven audio-visual approach based on the unit-selection algorithm [4] and photo-realistic image sequences [5].

So far though, there has hardly been any research or success in creating a German video-realistic audio-visual synthesis system. That is why we extended our framework for producing a data-driven video-realistic audio-visual "Talking Head" [15, 16], as the basis of our German multimodal Human-Computer-Interface, and defined a phoneme set based on BOSS [18] for the speech synthesis in German, as well as a viseme set for the corresponding visual segments. In order to create the appropriate mouth movement of the spoken utterances, video segments were selected according to the visemic transcription. The visemic transcription is important to audio-visual synthesis due to the fact that it reduces the amount of possible visual segments. This is because we are not able to see a difference in the visual realisation of phones like a /p/ and a /b/. Both are plosives and only the lip movement can be recognized during speech. For this reason we developed a viseme inventory with fifteen classes. The phoneme-viseme mapping for German audio-visual synthesis is already in use in our video-realistic audio-visual synthesis system.

According to Ratnarparkhi [13], we trained a Maximum-Entropy model for visemically transcription of corpora, providing feasible results.

This paper is organized as follows. In section two we describe our German phoneme and viseme set as well as the phoneme-viseme mapping. In section three we introduce our viseme inventory for German and in section four we show the visual representation of the phoneme realisation which is differentiable by humans. Also we list the results while training an automatic visemic transcription.

2. The Phoneme and Viseme Set

In order to drive talking heads including the usage of speech, creating an appropriate phoneme-viseme mapping is necessary. This approach is based upon the segmentation of the speech signal into discrete linguistic units, in this case phonemes, as well as upon the synthesis of facial lip motion by selecting appropriate units from our database which relate on the underlying visemic transcription of the input text according to the speech signal. Therefore, the transcription module of the audio-visual synthesis system converts the given graphemic input text into a phonemic and visemic representation during runtime. Many work is published phonemic transcription but less on visemic and non for German. Following we will introduce our phoneme set and

describe the phoneme viseme mapping to build our viseme inventory.

2.1. The German Phoneme Set

The phonemic transcription is used within the text preprocessing module in the speech synthesis part. Therefore, we rely on the transcription set being used in "BOSS II (DE)", the "Bonn Open Synthesis System II (for German)" [6, 18]. Its German transcriptions are quite similar to the SAMPA-DE-inventory. *Table 1* shows some examples of the according German phoneme set that we use to phonetically transcribe the text input which we want to synthesize within our audio-visual synthesis. For a detailed description of the BOSS II phoneme labels please see [6, 18].

Table 1: Examples of the German phoneme inventory

BOSS -DE	Example	BOSS-DE	Example
i:	Sie	p	Platz
y:	Grüße	b	Bär
E	Wetter	t	Tag
OY	Freude	C	Licht
@	Tage	N	Drang

2.2. The Phoneme-Viseme Mapping

According to the phonemic transcription inventory which is used in the speech synthesis part, a mapping of the phonemic units to a visemic set of symbols for representing the visual sequences of the later two-dimensional synthesis got developed next:

The first viseme class mapping consonants includes the two bilabial plosives /p/ and /b/. In comparable research for English synthesis systems, we found that /m/ was added to that class as well. We decided to assign an own class to the bilabial nasal though, because the place of articulation might be the same for the three phonemes, but the lips stay close for an /m/ in final position, which visually differentiates this sound from the other two released plosives. The four remaining plosives /t, d, k, g/ are put together too, although they differentiate in the place of closure: /t, d/ are produced by an alveolar, /k, g/ by a velar closure. But since this difference is produced within the mouth, it is not visually distinguishable.

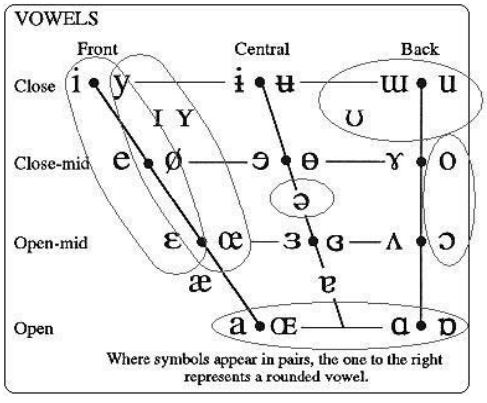


Figure 1: Vowel Chart and vocalic Viseme Classes

The next class is composed by /n, @n, l, @l/, two alveolar phonemes and their combination with /@/. The front of the tongue is shaped differently for /n/ and /l/, and the velum is lowered for /n/, but this again is not crucial to the visual observation.

The two labiodental fricatives /f/ and /v/ are combined to one viseme class, as well as the two alveolar fricatives /s/ and /z/. The next class includes postalveolar fricatives as well as the according affricates: / S, Z, tS, dZ/, but again, this difference, the additional plosives, is not visually perceptible.

The three fricatives /h/, /r/ and /x/ and the nasal /N/ build up another class, though having different places of articulation. But the articulators' approximation takes place in the mouth's back, while the lips stay open, so those different places will not alter the visual perception. Having the same place of articulation and just differing in the amount of approximation, within the mouth, the two phonemes /j/ and /C/ form the last of the consonantal viseme classes.

For the vowels' viseme classes, we combined the phonemes according to the height of the tongue's body and its front-back position, or rather according to their similarities, as illustrated in *Figure 1*. Consequently, we defined the following viseme classes: /i:, I, e:, E:, E/, /a:, a/, /o:, O/ and /u:, U/, and the neutral phonemes /@/ and /6/ build up one viseme class, as well as the four rounded vowels /y:, Y, 2:, 9/. Apparently, when we defined these six vocalic viseme classes, we did not differentiate between the tensed and lax vowels. This is because they hardly differ in their production the slightly different articulation of the pairs of tensed and lax vowels will not be visually recognized and distinguished.

3. The German Viseme Set

A viseme is a generic image of the mouth shape that can be used to describe a particular sound and represents the visual equivalent of a phoneme or unit of sound in spoken language, as illustrated in *Table 3*. As a symbol inventory for the presented fifteen viseme classes, the visual counterparts to the phonemic transcriptions, we agreed on using capitalized letters, in general reflecting the phonetic pronunciation of one classes' sound as *Table 2* shows.

Table 2: Viseme Inventory and their Phonemic Mapping

No.	Phoneme (BOSS-DE)	Viseme	Example
1	p, b	**P**	Pause, Bitte
2	t, d, k, g	**T**	Tonne, Dach, König, Gier
3	n, @n, l, @l	**N**	Nadel, raten, Liebe, Igel
4	m	**M**	Mutter
5	f, v	**F**	Finder, Vase
6	s, z	**S**	Fass, Sein
7	S, Z, tS, dZ	**Z**	Schar, Rage, Tscheche, Dschungel
8	h, r, x, N	**R**	Hase, Reden, Dach, Wange
9	j, C	**C**	Junge, Wicht
10	i:, I, e:, E:, E	**E**	Bier, Tisch, Weg, Räte, Menge
11	a:, a	**A**	Wagen, Watte
12	o:, O	**O**	Wolle, Wogen
13	u:, U	**U**	Buch, Runde
14	@, 6	**Q**	Bitte, Weiher
15	y:, Y, 2:, 9	**Y**	Tür, Mütter, Goethe, Götter

The diphthongs, such as /aU/ or /aI/, the combinations /vowel+6/, as well as the affricate /pf/ are expressed by the composition of the corresponding two viseme classes, for example [AU] for /aU/, [OE] for /OY/ or [EQ] for /I6/.

Examplary Phoneme-Viseme Transcription:
"Wetterauskunft" transcribed as

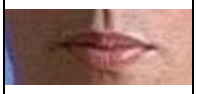
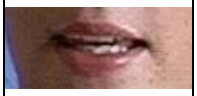
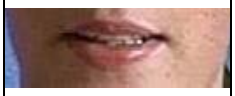
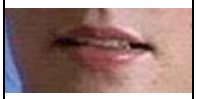
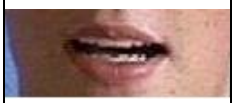
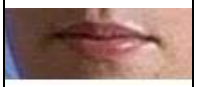
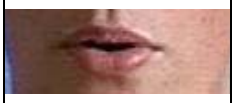
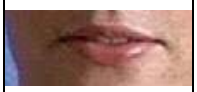
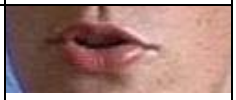
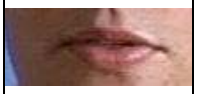
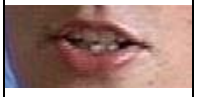
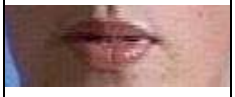
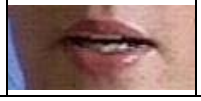
- Phonemically = v E t 6 ?aU s k U n f t
- Visemically = F E T Q AU S T U N F T

This viseme set is used in our audio-visual synthesis system where we have automatically transcribed our recorded speech corpora.

4. Visual representation of the German Viseme Inventory

The visual segments which we use in our audio-visual synthesis system, according to the discussed and listed fifteen basic viseme classes, are displayed in the following screenshots of mouth and lip positions in *Table 3*.

Table 3: Visual representation of the fifteen viseme classes

Vi-seme	Visual Representation	Vi-seme	Visual Representation
P		***C***	
T		***E***	
N		***A***	
M		***O***	
F		***U***	
S		***Q***	
Z		***Y***	
R			

4.1. Maximum Entropy Training for visemic transcription

We trained the model with the Maximum Entropy learning approach, according to Berger et al. [2], and used general iterative scaling Maximum Entropy Leraning for natural language disambiguation, introduced by Ratnarparkhi [14].

For our visemic transcription training part we produces a training file. The details about the resulting training are shown *Table 4*.

Table 4: Viseme Inventory and their Phonemic Mapping

Training Data	
read contexts	368
Number of Features	1524
Model: Threshold	1.0E-4
Maximum Iterations	100
Performance	
Log-likelihood	-1201.49
Performance	95.34%

This seems to be an appropriate outcome and a vantage point for deployment, further research or refinement.

5. Conclusions

In this paper we introduced a German viseme set based on the SAMPA-D phoneme inventory to be used within our data-driven audio-visual synthesis systems. This German viseme inventory can serve as a source and originator for modification, addressing different claims and needs, such as the viseme classes' utilization within audio-visual speech recognition systems or audio-visual speech synthesis systems . An automatic grapheme-to-viseme transcription was explored using the statistically motivated maximum-entropy approach, and the according results show an adequate performance and quality of the system. Hence, we already use this automatic visemic transcription in our current data-driven audio-visual synthesis system for an appropriate visual segment selection.

6. References

[1] Bailly, G., Bérar, M., Elisei, F., Odisi, M.: "Audiovisual Speech Synthesis". In: *International Journal of Speech Technology*, Vol.6. October 2003.

[2] Berger, A. L., Della Pietra, S. A., Della Pietra, V. J.: "A Maximum Entropy Approach to Natural Language Process". In: *Computational Linguistics*, Vol. 22. 1996.

[3] Beskow, J.: "Talking Heads - Models and Applications for Multimodal Speech Synthesis". PhD Thesis. Stockholm: June 2003.

[4] Black, A., Campbell, N.: "Optimizing selection of units from speech databases for concatenative synthesis". In: *Eurospeech*, Vol. 1. Madrid: 1995

[5] Bregler, C., Covell, M., Slaney, M.: "Video Rewrite: Driving Visual Speech with Audio". In: *Proc. SIGGRAPH, ACM SIGGRAPH*. July 1997.

[6] Breuer, S., Abresch, J., Wagner, P., Stöber, K., Bröggelwirth, J.: "Documentation for Bonn Open Synthesis System (BOSS) II". In: *Internal report, Institut für*

Kommunikationsforschung und Phonetik, Universität Bonn. Bonn: October 2001.

[7] Breuer, S., Abresch, J., Wagner, P., Stöber, K.: "BLF - ein Labelformat für die maschinelle Sprachsynthese mit BOSS II". In: *Hess, W., Stöber, K. (Hrsg.): Tagungsband Elektronische Sprachsignalverarbeitung ESSV'2001, Studientexte zur Sprachkommunikation*. Bonn: 2001.

[8] Cohen, M. M., Massaro, D. W.: "Modeling Coarticulation in Synthetic Visual Speech, Models and Techniques in Computer Animation". Springer Verlag, New York: 1993.

[9] Cohen, M. M., Walker, R. L., Massaro, D. W.: "Perception of synthetic visual speech". In: *Stroke, D. G., Hennecke, M. E. (Eds.)*: *Speech reading by humans and Machines.* Springer Verlag, New York: 1996.

[10] Karlsson I., Faulkner A., Salvi G.: "SYNFACE - a talking face telephone. The Eurospeech Special Event on "Spoken Language Technology in E-inclusion" ", In: *Proc of EuroSpeech*. Geneva: September 2003.

[11] Massaro, D. W.: "Perceiving talking faces: From speech perception to a behavioral principle". The MIT Press, Cambridge, MA: 1998.

[12] McGurk, H., MacDonald, J.: "Hearing lips and seeing voices," in: *Nature*, vol. *264*, 1976.

[13] Pandzic, J., Ostermann, J., Millen, D.: "User Evaluation: Synthetic talking faces for interactive services." In: *The Visual Computer*. Springer Verlag, New York: 1999.

[14] Ratnarparkhi, A.: "Maximum Entropy Models for Natural Language Ambiguity Resolution". PhD Dissertation. University of Pennsylvania: 1998.

[15] Weiss, C.: "A Framework for data-driven video-realistic audio-visual speech synthesis". In: *Proceedings of Fourth Int. Conf. on Language Resources and Evaluation*. Lisbon: May 2004.

[16] Weiss, C.: "Videorealistische audiovisuelle Synthese basierend auf Unit-Selection". In: *Kroschel, C.: Konferenz "Elektronische Sprachsignalverarbeitung",* Tagungsband 14. Karlsruhe: September 2003.

[17] Baldi: <http://cslu.cse.ogi.edu/toolkit/>

[18] BOSS: <http://www.ikp.uni-bonn.de/dt/forsch/phonetik/boss/index.html>

[19] SAMPA: <http://www.phon.ucl.ac.uk/home/sampa/german.htm>

[20] Synface: <http://www.speech.kth.se/synface/>