

**Phoneme-Viseme Mapping for German
Video-Realistic Audio-Visual-Speech-Synthesis
IKP - Working Paper NF 11**

Bianca Aschenberner, Christian Weiss

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

Institut für Kommunikationsforschung und Phonetik, Universität Bonn

Abstract

In this working paper we introduce a German viseme set which we already use in our data-driven audio-visual synthesis system. The viseme set 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. The transcription also serves as label 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 visemes in related literature and serves also as unit label in our data-driven video-realistic synthesis system.

A viseme could include several phonemes, since we can not for example visually differentiate between pronouncing a /b/ or a /p/. Thus, both sounds are bilabial plosives and consequently, have the same lip-based realisation. Therefore, we worked out an inventory of German viseme classes and developed a phoneme-viseme mapping for German that we currently use in our data-driven exemplary audio-visual synthesis system. The visemes are represented in a SAMPA-like labelling. For transcribing a large corpus, we trained a maximum entropy model for automatic visemic transcription and thus received feasible results.

1 Introduction

Audio-visual synthesis becomes more and more interesting to both research and industry, because the fusion of the auditory and the visual representation provides a human-machine interface which is 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. McGurk and McDonald [12] have shown that the perception of speech by a person not only depends on acoustic cues, but also on visual cues such as lip movements. In this manner Massaro [11] for example showed in his work that the facial gesture plays a major role in face-to-face-communication and in Human-Machine-Interaction by improving the intelligibility of a spoken utterance especially in noisy environments (also known as the cocktail party effect) or for hearing impaired people. Because of the importance of the face and speech in all human-to-human and human-to-avatar interactions, this research is considered essential in numerous areas such as speech-reading education, E-commerce, customer relations as well as health services. In order 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. Evaluations [3, 10] on this topic show the increase of intelligibility and understanding when lip-movement is added to synthesized utterances. Examples of computer animated “Talking Heads” which produce audio and lip synchronized speech can be found in various applications [12], as mentioned above. Two major approaches are currently deployed: the model-based approach of building a Talking Head such as Baldi and Synface [17, 20] and the data-driven audio-visual approach which is based on the well known 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. For this reason, a framework was developed [15, 16] to produce a data-driven video-realistic audio-visual “Talking Head”, which can be used as a multimodal Human-Computer-Interface. In order to create the appropriate mouth movement of the spoken utterance, the video segments are selected according to the visemic transcription. To that effect, we developed a phoneme-viseme mapping for applications of video-realistic audio-visual synthesis in German. This included the definition of our phoneme set, being based on BOSS [18], which we use for the speech synthesis in German as well as the definition of a viseme set for the corresponding visual segments. For the visemic

transcription we trained a Maximum-Entropy model, according to Ratnarparkhi [13], for large-scale corpora.

In the following, we want to present the process of defining the basic German phoneme inventory, integrating the additional visual information and evolving the according set of viseme classes. Therefore, we describe our phoneme set in section 2 that relies on standard German-SAMPA [19], though with some adaptations corresponding to the BOSS-Documentation [6] in order to fulfil our specific requirements. In section 3, we specify the phoneme-viseme mapping, and list the according inventory which we use for the visemic output of the audio-visual synthesis in section 4. Section 5 finally shows the corresponding visual representations in order to exemplify and prove our viseme inventory as well as the results of our training process.

2 German Phoneme Set

In order to drive talking heads including the usage of speech, we first have to create an appropriate phoneme-viseme mapping. This approach is based upon the segmentation of the speech signal into discrete linguistic units, in this case phonemes and visemes, as well as upon the synthesis of facial motion by selecting appropriate units from our database which relate on the underlying visemic transcription of the speech signal. A number of talking heads have been implemented following this approach. This visemic transcription is obtained through a high-level statistical learning approach (re: section 5). During runtime, the transcription module of the audio-visual synthesis system converts the given graphemic input text into a phonemic and visemic representation. The phonemic transcription is used within the text preprocessing module in the speech synthesis part. Therefore, we rely on the transcription set which is used in “BOSS II (DE)”, the “Bonn Open Synthesis System II (for German)” [6, 18]. Its German transcriptions are quite similar to the SAMPA-D-inventory. In *Table 1* on the next page, we show the according German phonemic inventory that we use to phonetically transcribe the text input which we want to synthesize within our audio-visual synthesis.

Vowels

Consonants

BOSS (DE)	Example		BOSS (DE)	Example
i:	Sie		p	Platz
I	Bitte		b	Bär
y:	Grüße		t	Tag
Y	Mütter		d	Daumen
e:	Weg		k	Kopf
E	Wetter		g	Größe
E:	Säle		f	Fahrer
2:	Goethe		v	Vase
9	Götter		s	Fass
@	Tage		z	Sonne
6	Lager		C	Licht
a:	Rat		x	Dach
a	Mann		m	Mann
o:	Woge		n / @n	Tannen
O	Wolle		N	Drang
u:	Buch		l / @l	Wolkenhimmel
U	Runde		j	Junge
aI	Weise		r	Runde
aU	Rauch		h	Hoch
OY	Freude		S	Schein
i:6	Tier		Z	Rage
<p>Like the latter transcription /i:6/, the other diphthongs as in “der”, “Bar”, “Turm”, etc. are transcribed equivalent,:</p> <p>/ vowel + 6 /</p>			tS	Tscheche
			dZ	Dschungel
			? (Glottal Stop)	Arbeit

Table 1: German phoneme set BOSS-SAMPA

3 Phoneme-Viseme Mapping

According to the listed phonemic transcription inventory which is used in the speech synthesis part, we now need to map the phonemic units to a visemic set of symbols which will represent the later visual sequence according to the spoken utterance. In the following, we will therefore display the considerations and steps we did in order create an appropriate mapping of the phonemic and visemic segments of the later two-dimensional synthesis.

To begin with the consonants, the first viseme class includes two phonemes, /p/ and /b/, both being bilabial plosives. In comparable research and literature 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.

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/ have different places of articulation, namely glottal, uvular and velar. But nevertheless, we put them into one viseme class. The articulators' approximation takes place in the back of the mouth, 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 viseme classes for the consonants.

For all of these classes, we ignored the remaining distinction, the possible time differences between the voiced and voiceless consonants, such as voice onset time. But these are that minimal that they are not crucial to the consonants differentiation within the visual observation. Besides, we did not assign the glottal stop to a viseme class, but its realisation is hardly displayed by any articulatory movement or time changes, and thus is insignificant and visually not noticeable.

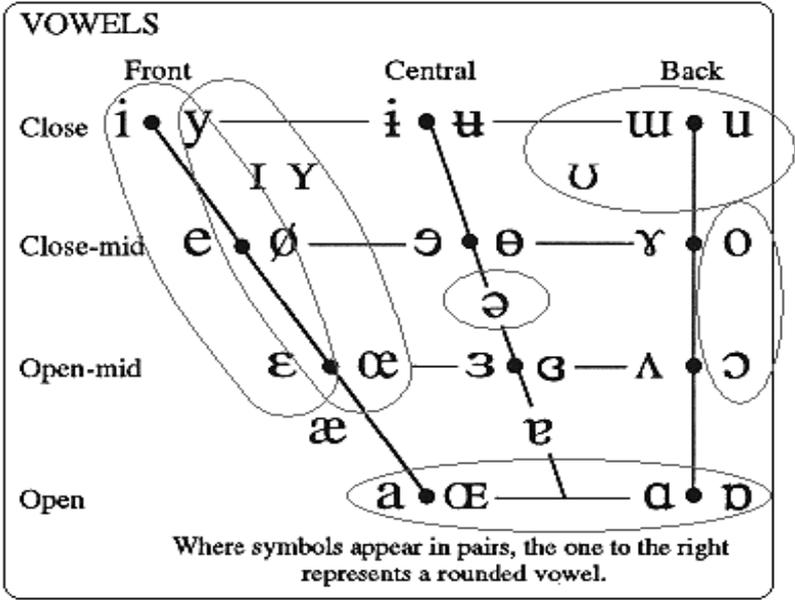


Figure 1: Vowel Chart

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/. Besides, 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 visemic classes for the vowels, 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.

4 German Viseme Set

A viseme is a generic image of the mouth shape that can be used to describe a particular sound. A viseme is the visual equivalent of a phoneme or unit of sound in spoken language, as illustrated in *Table 3*. Using visemes, especially the hearing-impaired can read sounds visually. As a symbol inventory for the presented viseme classes, the visual counterparts to the phonemic transcription, we agreed on using capitalized letters. In general, these letters reflect the phonetic pronunciation of one classes' sound. The corresponding viseme inventory is shown in *Table 2*:

No.	Phoneme (BOSS)	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	Schein, Garage, 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

Table 2: Phoneme-viseme mapping

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/, or [EQ] for /I6/. The only exception to these viseme combinations is the visemic representation of the diphthong /OY/. This phoneme combination is not mapped by the

visemic combination of [O] and [Y] which might be reasonable, but by [OE]. Thus, this latter combination [OE] for the diphthong /OY/ represents the according articulation and thus the visual output more adequately.

Phoneme-Viseme Transcription Example:

Wetterauskunft - v E t 6 ?aU s k U n f t - F E T Q A U S T U N F T

5 Visual representation of German Viseme Set

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 *Table 3* below.

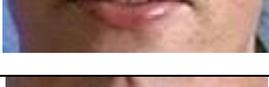
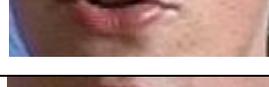
Visem	Visual Representation	Viseme	Visual Representation
P		C	
T		E	
N		A	
M		O	
F		U	
S		Q	
Z		Y	
R			

Table 3: Visual representation of visemes

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

Our input data consist of:

Read 368 contexts

Number of Features: 1524

Model: Threshold=1.0E-4

Model: Maximum Iterations=100

The current result performance is:

Log-likelihood: -1201.49, Performance: 95.34%

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

Conclusion

In this working 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. The latter 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. 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.

References

- [1] Bailly, G., Béjar, 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, Studentexte 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/>>