Journal of Computer Science and Information Technology
December 2017, Vol. 5, No. 2, pp. 1-9
ISSN: 2334-2366 (Print), 2334-2374 (Online)

Published by American Research Institute for Policy Development
DOI: 10.15640/jcsit.v5n2a1
URL: https://doi.org/10.15640/jcsit.v5n2a1

Foreign language visemes for use in lip-synching with computer-generated audio

John Whipple[1], Ruth Agada[2] & Jie Yan[3]

Abstract

There are several state-of-art for animating human motion, most of which involves the use of markers on the human and a tracker that estimates movement based on the position and orientation of these markers. In this paper, we discuss the different methods in which to extract human lip movement from video and map to the corresponding viseme of a foreign language for smooth animation of a 3D model. We discuss the use of Active shape model for obtaining lip movements, the use of established grapheme to phoneme methods and the commonality with the English phonemes, and how these are transferred onto a 3D human model

1. Introduction

Computerized text-to-speech synthesis with visual representation is a multifaceted problem domain requiring an understanding of a variety of subjects to arrive at a plausible solution. First, there is the step of converting the text into some form of audio. Basically, you take each word in the sentence and break it down into the basic sounds that make it up, called phonemes. Then you string all those phonemes for each word together with all the phonemes for the rest of the words in the sentence and you have an audio representation of that sentence. The process is not actually that straightforward however, because the English language contains many homographs and heteronyms that are problematic for the translator. Homographs are words that are spelled the same but have different meanings. These words inject ambiguity into the parsing of a sentence and make it difficult to find out exactly what the part of speech of the word is, and different parts of speech can alter the flow of a sentence and affect the meaning of other parts of that sentence as well, including their pronunciation, which is exactly what we are trying to produce, based solely on textual information. Heteronyms are a problem that further complicates that situation by taking the homograph, the word with the same spelling but different meaning, and then adding the fact that it is also pronounced differently. Without the existence of heteronyms, we could simply look up each word in a database and extract a string of phonemes with a one-to-one mapping. But we cannot do that because heteronyms do exist. Because of these heteronyms, we must analyze how that word is used in the sentence to determine which version of that heteronym we are dealing with so that we can generate the correct audio to represent it(Bear, Harvey, Theobald, & Lan, 2014; Bozkurt, Erdem, Erzin, Erdem, & Ozkan, 2007; Dave & Patel, 2014; Yu, Garrod, & Schyns, 2012).

2. Related work

Processing text to derive phonemes and then taking those phonemes to derive visemes is a defined process, and if you are starting with a predefined set of written text that you want to generate, that process suits your problem domain just fine. What if you had English source material and you wanted to present it in another language? This is a situation that occurs more frequently as the internet continues to shrink the world we live in.

[1]Bowie State University. Bowie, MD, USA,
[2]Bowie State University. Bowie, MD, USA,
[3]Bowie State University. Bowie, MD, USA,

If your only concern were text-to-speech synthesis with a visual representation, you would first need to translate that English text into text of the target foreign language, say Arabic, and then put that Arabic text through the same process that you used for the English text. You would first analyze the text to produce phonemes, and with those phonemes you could determine the resultant visemes that are necessary to display that performance visually.

Another problem facing text-to-speech synthesizers is that of stringing words together in a sentence. A single word by itself could be pronounced one way, but when there is another word trailing it in that sentence, the phonemes that exist on the boundary between those words could be altered, or one of them could possibly be completely erased because they are both merging together to form one single phoneme. Once you solve these problems however, you have some speech. It may not be a perfectly natural and human-sounding representation of that text, but it should be close enough to be intelligible. The other half of that two-sided coin is the visual representation of the speech.

Once the text has been broken down into its individual phonemes, each phoneme is mapped to a visual depiction of the mouth as it is producing that sound, called a viseme. Human beings have an infinite number of producible phonemes and visemes, but in practice, like how analog music is digitized, many similar phonemes that are almost indistinguishable from one another can be treated as one logical phoneme, and thus the overall number of phonemes that are actively used in each language is effectively constrained. Different languages have different numbers of phonemes. For example, English has around 48 phonemes that map to 15 possible visemes while Lithuanian has 58 phonemes and 16 visemes (Mažonavičiūtė, I. & Baušys, 2009). Japanese has 24 phonemes(Sawai, 1991). Arabic has 29 phonemes and 20 visemes(Chelali & Djeradi, 2011). Indonesian has 49 phonemes mapping to 12 visemes (Setyati et al., 2015). Khwe, a Khoisan tribal clicking language from Africa, has 70 phonemic consonants (including 35 clicks), and 25 vowel phonemes(Dixon, 2006). Just as there were complications with phonemes when words are strung together, the mapping of phonemes to visemes can be altered when certain phonemes are used in sequence. As such, Mattheyses, Latacz, and Verhelst posit that the true mapping of phonemes to visemes is not many-to-one, but rather many-to-many(2013).

America's film industry produces films that are presented all around the world to audiences who do not speak or understand English. Sometimes it is acceptable to present the movie with the original English audio and display subtitles across the bottom of the screen for the audience to read while they are watching the movie so they can follow what is going on in the story. That is a relatively easy proposition for the moviemakers; they simply contract out the translation to an in-house or outside body, who understands English and the target language, then pay to have them watch the movie and transcribe the spoken word into the target subtitles that will be used for the movie. Unfortunately for moviemakers, some cultures are less accepting of subtitled movies and they are used to consuming their media with dubbed audio in their own language, in fact they demand that their movies be made available to them in their native language (Garrido et al., 2015).When a movie is played back with its original audio, all the phonemes will match up in time with the visemes on screen, but once you swap out the audio with that of a new language, the new phonemes being played back will not synchronize with the original visemes being displayed on the screen.In the past, efforts were made to tailor the foreign language script to match up with the picture by choosing words that would at least approximate the original language visemes on display. Due to the many-to-one mapping of phonemes to visemes, these translating scriptwriters were granted a certain amount of leeway in their choice of wording.

Modern digital technological advances have made it possible to do the opposite. Garrido et. al.(2015)has developed a system that records a mapping of mouth shapes to time stamps in the original movie, and then processes the foreign language dubbed audio to obtain phonemes and determine the desired visemes that should be displayed at certain time stamps within the dubbed movie.Then they can alter the visual depiction of the performance in such a way that it will be a closer match to the dubbed audio, without having to alter the speech too much and make it seem stilted and unnatural(Garrido et al., 2015).

Another use for these phoneme-viseme mappings is in video games. Usually, a video game is written in one language, in such a way that all the text that will be visible in the game is contained in one file, or text database. When the publisher of that video game wishes to sell that game in another region, he will send just that file off to the localization department, or even a third-party localization service, and they will translate the text and send it back. With that translated text database, the game will be able to display the text in the appropriate language for the region in which it is being played, based on the language configuration of the host system.

Back in the era of 16-bit game consoles like the Sega Genesis/Mega Drive and the Super Nintendo/Super Famicon, that was all that was needed. Those systems were not capable of playing high fidelity voice audio, and even if they were, the cartridge media that held the games was far too limited in capacity to contain it, so all the dialogue was text-only. The video game systems of today are equipped hard disk drives and CD-ROM drives that can hold vast amounts of voice and video data, and the audio and video processors are easily capable of making use of it all. But still, many games limit themselves to text-only presentation, because the costs of localizing anything more than that would be prohibitive.

Larger companies have the financial resource to localize even their audio and video data, but they are taking a big risk. It takes a lot of time and money to accomplish all that localization, and if sales do not perform well in the target market, they stand to lose a large sum of money.

However, equipped with an effective text-to-speech system, all that translated text could easily be extrapolated to present audio and visual performance to the user as well, with little added cost to the publisher of the game. Such systems are frequently referred to as "talking heads" in the scholar community(Mažonavičiūtė, I. & Baušys, 2009; Raheem Ali, Sulong, & Kolivand, 2015).That is an apt description of what the head is doing, and for the purposes of that research I suppose it is all they are concerned with, but the technology can be applied to much more than just a talking head. Every character in a video game is essentially a talking head that also just happens to have a body attached to it so it can do other things and make the game more fun, but the talking head is where the language normally comes out of a person, so that is the portion of the character we are concerned with when talking about speech synthesis and its visual representation.

3. Speech corpus

From Czyzewskiet al. (2017), the Czech audio-visual database UWB-07-ICAVR (Impaired Condition Audio Visual speech Recognition) (Trojanová, Hrúz, Campr, & Zelezny, 2008) is focused on extending existing databases by introducing variable illumination, similar to VALID. The database consists of recordings of 10000 continuous utterances (200 per speaker; 50 shared, 150 unique) taken from 50 speakers (25 male, 25 female). Speakers were recorded using two microphones and two cameras (one high-quality camera, one webcam). Six types of illumination were used during every recording. The UWB-07-ICAVR database is intended for audio-visual speech recognition research. To aid it, the authors supplemented the recorded video files with visual labels, specifying regions of interest (a bounding box around mouth and lip area), and they transcribed the pronunciation of sentences into text files.

Audiovisual Polish speech corpus (AGH AV Corpus) (AGH University of Science and Technology 2014) is another example of an AVSR database built for Polish language. It is the largest audiovisual corpus of Polish speech (Igras M., Ziółko B., 2012; Jadczyk & Zi, 2015) as reported by Czyzewski et al. (2017). The authors of this study evaluate the performance of a system built of acoustic and visual features and Dynamic Bayesian Network (DBN) models. The acoustic part of the AGH AV corpus is more thoroughly presented and evaluated in the paper by the team of the AGH University of Science and Technology(Żelasko, Ziółko, Jadczyk, & Skurzok, 2016). Besides the audiovisual corpus, presented in Table 1, authors developed various versions of acoustic corpora featuring the large number of unique speakers, which amounts to 166. This results in over 25 hours of recordings, consisting of a variety of speech scenarios, including text reading, issuing commands, telephonic speech, phonetically balanced 4.5 hourssub corpus recorded in an anechoic chamber, etc. (Czyzewski et al., 2017).

Table 1. Comparison of existing databases from Czyzewski et al. (2017)

Database	Year	Number of speakers	Resolution	Fps	Language material	Additional features
TULIPS1	1995	12	100×75	30 fps	numerals 1–4	no
DAVID	1996	123	640×480	30 fps	numerals, alphabet, nonsense utterances	varying background
M2VTS	1997	37	286×350	25 fps	isolated numerals 0–9	head rotations, glasses, hats
XM2VTS	1999	295	720×576	25 fps	3 sentences (numerals and words)	head rotations, glasses, hats
CUAVE	2002	30	720×480	29.97 fps	isolated or connected numerals (7000 utterances total)	simultaneous speech
BANCA	2003	52	720×576	25 fps	numerals, name, date of birth and address	controlled, degraded and adverse conditions, impostor recordings
AVICAR	2004	84	360×240	29.97 fps	Isolated numerals and letters, phone numbers, TIMIT sentences	automotive noise, microphone, and camera array
VALID	2005	106	720×576	25 fps	same as XM2VTS	varying illumination and noise
GRID	2005	34	720×576	25 fps	1000 command-like sentences	no
DXM2VTS	2008	295	720×576	25 fps	same as XM2VTS	varying background, video distortions
VIDTIMIT	2008	43	512×384	25 fps	10 TIMIT sentences	office noise and zoom
UWB-07-iCAV	2008	50	720×576	max 50 fps	continuous Czech utterances	varying illumination and quality
IV2	2008	300	780×576 max	25 fps	15 French sentences	stereo frontal and profile views, iris images, 3D scanner data, head pose and illumination variations
WAPUSK20	2010	20	640×480	48 fps	100 GRID sentences	stereoscopic camera, office noise
BL	2011	17	640×480	30 fps	238 French sentences	depth camera, highlighted lips
UNMC-VIER	2011	123	708×640 max	29 fps	12 XM2VTS sentences	varying quality, speech tempo, expressions, illumination, head poses
MOBIO	2012	152	640×480	16-30 fps	32 questions	recorded on mobile devices, varying head pose and illumination
AGH AV Corpus	2014	20	1920×1080	25/50 fps	Isolated words, numerals	Polish language, audio: 16 bit/44.1 kHz, h.264 video codec
MODALITY	2015	35	1920×1080	100 fps	168 commands (isolated, sentences)	stereo camera, varying noise, microphone array, word SNR, additional depth camera

4. Methodology

4.1. Active Shape model

An active shape model (ASM)(Tresadern, Ionita, & Cootes, 2011; van Ginneken, Frangi, Staal, ter Haar Romeny, & Viergever, 2002a)consists of a Point Distribution Model (PDM) aiming to learn the variations of valid shapes, and a set of flexible models capturing the grey-levels around a set of landmark feature points. Active shape models are based on many implicit but crucial assumptions: (i) the shape of the object of interest can be defined by a relatively small set of explicit view models, (ii) the grey levels around a landmark point are consistent for all the views of the object and can be used to find correspondences between these views and, (iii) the shapes at different views vary linearly. These assumptions are valid when the variations allowed are well constrained.

From van Ginneken et al. (2002), an object is described by points, referred to as landmark points. The landmark points are (manually) determined in aset of training images. From these collections of landmark points, a point distribution model is constructed as follows. The landmark points $\{x_1, y_1\}, \dots \{x_n, y_n\}$ are stacked in shapevectors

$$\mathrm{X} = (x_1, y_1, \dots, x_n, y_n)^T \quad (1)$$

Principal component analysis (PCA) is applied to the shape vectors by computing the mean shape

$$\bar{\mathrm{x}} = \frac{1}{\mathrm{s}} \sum_{\mathrm{i}=1}^{\mathrm{s}} \mathrm{x_i} \quad (2)$$

the covariance

$$\mathrm{S} = \frac{1}{\mathrm{S}-1} \sum_{\mathrm{i}=1}^{\mathrm{s}} (\mathrm{x_i} - \bar{\mathrm{x}})\,(\mathrm{x_i} - \bar{\mathrm{x}})^{\mathrm{T}} \quad (3)$$

and the eigen system of the covariance matrix. The eigen vectors corresponding to the *t* largest eigen values λ_i are retained in a $\Phi = (\phi_1 | \phi_2 | \dots \phi_t)$. A shape can now be approximated

$$x \approx \bar{x} + \Phi b \quad (4)$$

b is a vector of t elements containing the model parameters, computed by

$$b = \Phi^T(x - \bar{x}) \quad (5)$$

when fitting the model to a set of points, the value of **b** is constrained to lie within the range $\pm m\sqrt{\lambda_i}$, where m has a value between two and three.

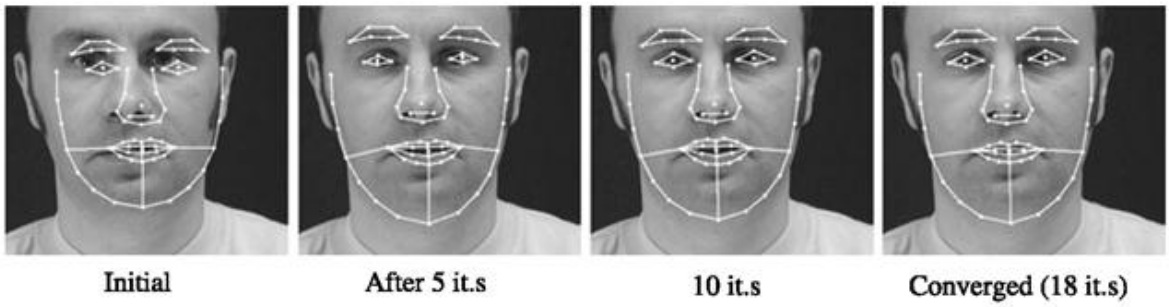

Fig.1. Search using Active Shape Model of a face.

4.2. Grapheme-to-phoneme

From Bisani and Ney (2008), the Automatic grapheme-to-phoneme conversion was first considered in the context of text-to-speech (TTS) applications. After normalization (expanding abbreviation, numerals, etc.) the input text needs to be converted to a sequence of phonemes which is then used to control a speech synthesizer. The simplest technique is dictionary look-up. While eff ective, it has serious limitations: Making a pronunciation dictionary of significant size (over 100,000 entries) by hand is tedious and therefore costly. Also, the storage requirements of such a database can be problematic for embedded or mobile devices. More importantly, a finite dictionary will always have limited coverage, while TTS systems are often expected to handle arbitrary words.

To overcome the limitations of simple dictionary look-up, rule-based conversion systems were developed. These can typically be formulated in the framework of finite-state automata(Kaplan & Kay, 1994). Often rule-based G2P systems also incorporate a dictionary as an exception list. While rule-based systems provide good (or even complete) coverage they have two drawbacks: Firstly, designing the rules is hard and requires specific linguistic skills. Secondly, natural languages frequently exhibit irregularities, which need to be captured by exception rules or exception lists. The interdependence between rules can be quite complex, so rule designers must cross-check if the outcome of applying the rules is correct in all cases. This makes development and maintenance of rule systems very tedious in practice. Moreover, a rule-based G2P system is still likely to make mistakes when presented with an exceptional word, not considered by the rule designer(Bisani & Ney, 2008).

In contrast to the knowledge-based approach outlined above, the data-driven approach to grapheme-to-phoneme conversion is based on the idea that given enough examples it should be possible to predict the pronunciation of unseen words purely by analogy. The benefit of the data-driven approach is that it trades the intellectually challenging task of designing pronunciation rules, for the much simpler one of providing example pronunciations. For native speakers, it is much easier to judge the correctness of a pronunciation or to write down the pronunciation of a specific word, than to formulate general spelling rules. The crucial question in data-driven G2P is how analogy should be implemented algorithmically. Starting with the work of Sejnowski and Rosenberg (1987), various machine learning techniques have been applied to this problem in the past. Before we try to give an overview in the following, we note that there are two partly competing goals in data-driven G2P, namely lexicon compression and generalization. Lexicon compression aims to minimize the storage (and computational) requirements by minimizing the error on seen data using a compact model. Generalization aims to overcome the limited coverage of a given dictionary by minimizing error on unseen data(Bisani & Ney, 2008).

It is worth noting that the pronunciations used to train a data-driven G2P model ought to exemplify the pronunciation rules of the language. This is contrary to the exception list used by rule-based systems which only need to cover the atypical pronunciations.

Training a model using only words with exceptional pronunciations would clearly defy any analogy-based approach. In practice, available pronunciation dictionaries which typically cover the most frequent words of the language are often used to train data-driven G2P models. While such dictionaries usually do contain atypical words, the patterns found in the more frequent, exemplary words will normally prevail. In fact, the data-driven approach mitigates the distinction between rules and exceptions. Ultimately, training data should be representative of the application domain (Bisani & Ney, 2008).

Table 2. The grapheme and phoneme inventories of the recognizers from Stuker and Schultz (2004)

Graphemes	Phonemes	Graphemes	Phonemes
a	a	u	u
b	b	f	f
v	w	h	h
g	g	c	ts
d	d	q	tscH
e	ye	x	sch
¨e	yo	w	schTsch
	jscH	□	Q
z	z	y	i2
i	i	~	
i$	j		e
k	k		yu
l	l		ya
m	m		b#
n	n		d#
o	o		jscH#
p	p		m#
r	r		n#
s	s		p#
t	t		r#
			s#
			sch#
			tscH#
			w#
			z#

4.3. Viseme mapping

In this section, we describe the decision tree-based viseme clustering methods first proposed in (Galanes, Unverferth, Arslan, & Talkin, 1998; Rademan & Niesler, 2015), and subsequently expanded to many-to-many phoneme-to-viseme mappings in(Mattheyses et al., 2013; Rademan & Niesler, 2015). Both contributions discuss the application of regression trees to the grouping of static visemes. Clusters of static visemes are split by querying their phonetic context or properties. Since the decision tree algorithms test more than one attribute when attempting to split a group of visemes in a leaf node, they can be classified as multivariate CART algorithms(Loh, 2011; Quinlan, 1993; Rademan & Niesler, 2015; Witten, Frank, & Hall, 2011).

The decision tree described in (Mattheyses et al., 2013) applies all possible phonetic context questions to the static visemes grouped in a decision tree's leaf node. The algorithm then measures how homogeneous the resulting child nodes are. We the apply the active shape model for automatic markerless facial tracking, generating the parameters that numerically describe the static visemes.

Equation 6 is applied to each phoneme instance pi in a leaf node, where $d(p_i,p_j)$ is the Mahalanobis distance between a point p_i and $p_j$and a distribution D and N is the number of phonemes in the node. The smallest value μ_{best} and variance σ best are then selected. Equation 7 is then used to determine the subset impurity I_Z, in which λ is a scaling factor. This procedure is repeated to find the question whose subset best minimizes the impurity of the ASM parameters in the child nodes(Rademan & Niesler, 2015).

$$\mu_i = \frac{\sum_{j=1}^{N} d(p_i, p_j)}{N-1} \quad (6)$$

$$I_Z = N \times (\mu_{best} - \lambda \times \sigma_{best}) \quad (7)$$

5. Conclusion and Future work

Visemes are also useful in the process of lip-reading. The jwasi et al.,(2008) developed a method for analyzing video sequences to determine the location of the lips of the speaker and what visemes are being formed. Because of the many-to-one phoneme to viseme mapping relationship, lip-reading systems encounter a new difficulty in that they are starting with a viseme and are trying to map it to one of many possible phonemes.

Goldschen (1996), Bear(2014), and Capelletta (2012;2011) have researched which mapping system produces the best results, but there will always be some ambiguity due to the nature of the problem. Since different languages contain different viseme sets, you could feasibly analyze the visemes that a person produces in a conversation and narrow down adequately synthesis realistic animated lip sync of that language.

In this paper, we reviewed the established methods for modeling lip movement in video to isolate viseme for foreign languages using active shape model. In establishing the visual tracking method, we reviewed the current body of works in terms of the different studies and research work performed using each of these methods for converting

Fig.2. Sample visemes and phoneme morph targets.

grapheme to phoneme. We considered the hybrid methods implemented using several of the established methods of animated lip synchronization.

References

Bear, H. L., Harvey, R. W., Theobald, B.-J., & Lan, Y. (2014). Which Phoneme-to-Viseme Maps Best Improve Visual-Only Computer Lip-Reading? In G. Bebis, R. Boyle, B. Parvin, D. Koracin, R. McMahan, J. Jerald, … M. Carlson (Eds.), *Advances in Visual Computing: 10th International Symposium, ISVC 2014, Las Vegas, NV, USA, December 8-10, 2014, Proceedings, Part II* (pp. 230–239). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-14364-4_22

Bisani, M., & Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, *50*(5), 434–451. https://doi.org/10.1016/j.specom.2008.01.002

Bozkurt, E., Erdem, C. E., Erzin, E., Erdem, T., & Ozkan, M. (2007). Comparison of Phoneme and Viseme Based Acoustic Units for Speech Driven Realistic lip Animation. In *2007 3DTV Conference* (pp. 1–4). IEEE. https://doi.org/10.1109/3DTV.2007.4379417

Cappelletta, L., & Harte, N. (2011). Viseme definitions comparison for visual-only speech recognition. *European Signal Processing Conference*, (Eusipco), 2109–2113.

Cappelletta, L., & Harte, N. (2012). Phoneme-to-Viseme Mapping for Visual Speech Recognition. *1st International Conference on Pattern Recognition Applications and Methods (ICPRAM) Volume 2*, 322–329.

Chelali, F. Z., & Djeradi, A. (2011). Primary Research on Arabic Visemes, Analysis in Space and Frequency Domain. *Int. J. Mob. Comput. Multimed. Commun.*, *3*(4), 1–19. https://doi.org/10.4018/jmcmc.2011100101

Czyzewski, A., Kostek, B., Bratoszewski, P., Kotus, J., & Szykulski, M. (2017). An audio-visual corpus for multimodal automatic speech recognition. *Journal of Intelligent Information Systems*, 1–26. https://doi.org/10.1007/s10844-016-0438-z

Dave, M. N., & Patel, N. M. (2014). Phoneme and Viseme based Approach for Lip Synchronization. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, *7*(3), 385–394. Retrieved from http://dx.doi.org/10.14257/ijsip.2014.7.3.31

Dixon, R. M. (2006). Serial verb constructions: Conspectus and coda. In A. Y. Aikhenvald & R. M. W. Dixon (Eds.), *Serial verb constructions: A cross-linguistic typology* (pp. 338–350). New York: Oxford University Press.

Galanes, F. M., Unverferth, J., Arslan, L., & Talkin, D. (1998). Generation of lip-synched synthetic faces from phonetically clustered face movement data. In *Proc. International Conference on Auditory-Visual Speech Processing* (pp. 191–194). https://doi.org/10.1.1.398.9204

Garrido, P., Valgaerts, L., Sarmadi, H., Steiner, I., Varanasi, K., P?rez, P., & Theobalt, C. (2015). VDub: Modifying Face Video of Actors for Plausible Visual Alignment to a Dubbed Audio Track. *Computer Graphics Forum*, *34*(2), 193–204. https://doi.org/10.1111/cgf.12552

Goldschen, A. J., Garcia, O. N., & Petajan, E. D. (1996). Rationale for Phoneme-Viseme Mapping and Feature Selection in Visual Speech Recognition. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by Humans and Machines: Models, Systems, and Applications* (pp. 505–515). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-13015-5_39

Igras M., Ziółko B., J. T. (2012). Audiovisual database of Polish speech recordings. *Studia Informatica*, *33*(2B), 163–172.

Jadczyk, T., & Zi, M. (2015). Audio-Visual Speech Processing System for Polish with Dynamic Bayesian Network Models. In *Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science (EECSS 2015)* (pp. 1–7). Barcelona, Spain.

Kaplan, R. M., & Kay, M. (1994). *Regular Models of Phonological Rule Systems*. *Computational Linguistics*. Retrieved from http://portal.acm.org/citation.cfm?id=204917&dl=GUIDE%5Cnhttp://portal.acm.org/citation.cfm?id=204917%7B&%7Ddl=GUIDE

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23. https://doi.org/10.1002/widm.8

Mattheyses, W., Latacz, L., & Verhelst, W. (2013). Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis. *Speech Communication*, *55*(7–8), 857–876. https://doi.org/10.1016/j.specom.2013.02.005

Mažonavičiūtė, I., & Baušys, R. (2009). English talking head adaptation for lithuanian speech animation. *Information Technology and Control*, *38*(3), 217–224.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Rademan, C. F., & Niesler, T. (2015). Improved Visual Speech Synthesis using Dynamic Viseme k -means Clustering and Decision Trees. *Facial Analysis, Animation and Auditory-Visual Speech Processing (FAAVSP)*, 169–174. Retrieved from http://www.isca-speech.org/archive/avsp15/av15_169.html

Raheem Ali, I., Sulong, G., & Kolivand, H. (2015). Realistic Lip Syncing for Virtual Character Using Common Viseme Set. *Computer and Information Science*, *8*(3), 71–82. https://doi.org/10.5539/cis.v8n3p71

Sawai, H. (1991). TDNN-LR continuous speech recognition system using adaptive incremental TDNN training. *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, 53–56 vol.1. https://doi.org/10.1109/ICASSP.1991.150276

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, *1*, 145–168. https://doi.org/10.1016/0004-3702(89)90017-9

Setyati, E., Sumpeno, S., Purnomo, M. H., Mikami, K., Kakimoto, M., & Kondo, K. (2015). Phoneme-viseme mapping for Indonesian language based on blend shape animation. *IAENG International Journal of Computer Science*, *42*(3), 1–12.

Stuker, S., & Schultz, T. (2004). A grapheme based speech recognition system for Russian. In *Proc. SPECOM* (pp. 297–303).

Thejaswi, N. S., & Sengupta, S. (2008). Lip Localization and Viseme Recognition from Video Sequences. *Fourteenth National Conference on Communications*, 456–460.

Tresadern, P. a., Ionita, M. C., & Cootes, T. F. (2011). Real-time facial feature tracking on a mobile device. *International Journal of Computer Vision*, *96*(3), 280–289. https://doi.org/10.1007/s11263-011-0464-9

Trojanová, J., Hrúz, M., Campr, P., & Zelezny, M. (2008). Design and Recording of Czech Audio-Visual Database with Impaired Conditions for Continuous Speech Recognition. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 1239–1243. Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/

van Ginneken, B., Frangi, A. F., Staal, J. J., ter Haar Romeny, B. M., & Viergever, M. A. (2002a). Active shape model segmentation with optimal features. *IEEE Transactions on Medical Imaging*, *21*(8), 924–933. https://doi.org/10.1109/TMI.2002.803121

van Ginneken, B., Frangi, A. F., Staal, J. J., ter Haar Romeny, B. M., & Viergever, M. A. (2002b). Active shape model segmentation with optimal features. *IEEE Transactions on Medical Imaging*, *21*(8), 924–933. https://doi.org/10.1109/TMI.2002.803121

Witten, I. H., Frank, E., & Hall, M. a. (2011). *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook). Complementary literature None*. https://doi.org/0120884070, 9780120884070

Yu, H., Garrod, O. G. B., & Schyns, P. G. (2012). Perception-driven facial expression synthesis. *Computers and Graphics (Pergamon)*, *36*(3), 152–162. https://doi.org/10.1016/j.cag.2011.12.002

Żelasko, P., Ziółko, B., Jadczyk, T., & Skurzok, D. (2016). AGH corpus of Polish speech. *Language Resources and Evaluation*, *50*(3), 585–601. https://doi.org/10.1007/s10579-015-9302-y