Viseme Comparison Based on Phonetic Cues for Varying Speech Accents

Chitralekha Bhat, Sunil Kopparapu

TCS Innovation Labs, Mumbai

bhat.chitralekha@tcs.com, sunilkumar.kopparapu@tcs.com

10.21437/Interspeech.2015-676

Abstract

Human interaction through speech is a multisensory activity, wherein the spoken audio is perceived using both auditory and visual cues. However, in the absence of auditory stimulus, speech content can be perceived through lip reading, using the dynamics of the social context. In our earlier work [1], we had presented a tool enabling hearing impaired to understand spoken speech in videos, through lip reading. During evaluation it was found that a hearing impaired person, trained to lip read Indian English was unable to lip read speech in other accents of English. We hypothesize that this difficulty can be attributed to a difference in viseme formation arising from underlying phonetic characteristics. In this paper, we present a comparison between auditory and visual space for the same speech utterance in English, as spoken by an Indian and a Croatian national. Results show a clear correlation between distances in the visual and auditory domain at viseme level. We then evaluate the feasibility of building visual subtitles through viseme adaptation from unknown accent to known accent.

Index Terms: visual speech, cross modal comparison, viseme adaptation from known accent

1. Introduction

Speech is the most important signal for human-human interaction, the perception of which is not limited to signals from the auditory domain alone. During a face-to-face interaction, the percept is formed when the auditory and visual systems process the cues presented to them in a synchronous fashion. Studies report that perception of speech is faster when visuals are available [2, 3] indicating a strong correlation between visual and auditory domains when speech perception is considered. Influence of visual cues on interpretation of various aspects of speech such as prosody, speaking rate, coarticulation, lexical tone, emotions etc. have been explored. In their work [4], authors examined the contribution of visual and auditory cues to prosody perception and found that visual cues alone were sufficient for this task. In their further work [5], they specifically focus on the correlation between head and eyebrow movement and different prosodic conditions. Variations in speaking rate are more prominent in the visual domain than in the acoustic domain [6], the authors attribute this to the effect of coarticulation. Researchers have found that visual only (VO) cues can be used to distinguish and identify tones, by speakers of tonal languages like Thai and Mandarin, with non-native speakers actively seeking visual information for tone discrimination [7]. Several of their studies are focussed towards understanding the influence of visual cues in tone perception. Classification of emotions was achieved to a good extent through visual only features using Principal component analysis and speaker face normalization[8]. In [9], authors investigated the usefulness of visual cues for emotion identification and found a strong relationship between audio and video modalities for emotion perception. In recent work, researchers have been able to classify native/non-native speakers based only on visual cues [10], using bag of dynamic features, wherein Discrete Cosine Transform (DCT) features gave the best classification result, indicating a possibility of accent classification based on visual-only features. All of the above studies contribute to either automatic recognition of audio-visual signals or synthesis of visual representation of aspects of speech, for automatic creation of speech animation. Not much visual-speech related work has been done from the perspective of assistive technology for hearing impaired, which demands high accuracy of the visual representation of speech. A clear understanding of the contribution of visual and auditory cues on speech perception is imperative for building assistive tools. Literature suggests that information is better acquired when auditory and visual cross modal acquisition process is used. In [11], authors have explored the influence of auditory priming on lip reading performance of individuals, results suggested that phonological representations were shared across auditory and visual processing. Studies for understanding auditory-visual nature of speech are being conducted at the grass-root levels through developmental studies [12]. Some studies also reported that visual speech does not necessarily depend on the auditory cortex region being activated [13], implying that a fair amount of information regarding the sound content can be derived from the visual space alone. In [14] research directions towards understanding natural statistics of audio visual speech have revealed robust correlations and correspondence between the area of the mouth opening (inter-lip distance) and speech, both in temporal and frequency domain (first and second formants) as well as a good degree of correlation between the syllable timescale and facial movements.Researchers also reported consistency in results across two languages and different speech contexts.

In our earlier work [1], we had built a tool that was envisioned as a visual aid for the hearing impaired to enable access to videos. The tool essentially produces lip movement sequences for a selected video, such that it is in time synchronization with the speech of the video. The lip movement sequence is then superimposed onto and displayed along with the original video. This visual equivalent of closed captioning was termed visual subtitles. During user evaluation of this tool, it was found that the users were unable to perceive visual subtitles for non-Indian accents of English. We hypothesise that the difficulty is due to the difference in the visual domain and can be attributed to the difference in the auditory domain. The objective of this paper is to explore the correlation between phonetic cues in the auditory domain and viseme formation in the visual domain for the same speech utterance when spoken by in two different accents. We further explore if this difference can be inferred or predicted from the phonetic domain alone, by building a mapping between two accents (say A1 and A2) so that a person

trained to lip read in accent A1 can still lip read speech in accent A2.

The rest of the paper is organized as follows. Section 2 describes the technique used to compare the visemes for the same speech sound as spoken by natives of two different countries and is the main contribution of this paper, Section 3 discusses the experimental results, and we conclude in Section 4.

2. Computation of visual and phonetic distances

In this paper, we verify if there exists a difference in the visual representation (viseme) of the same speech sound when uttered by individuals from different nationalities and hence spoken in different accents. This exercise is envisioned to give insights into the influence on the formation of visemes for varying accents in speech for the same language, namely English. There does not exist a standardised definition of visemes in literature. In [15], authors describe two practical definitions and the method to arrive at them. The viseme classification used in this work is based on linguistic knowledge and is derived such that each viseme maps to a set of phonemes that have the same visual appearance, giving a one-to-many mapping of visemes to phonemes as shown in Table 1

Viseme	Phonemes	Viseme	Phonemes
0	sil	12	ay
1	E	13	h
2	a	14	r
3	A	15	l
4	O;au	16	s;z
5	ey	17	S;c;zh;j
6	er	18	t;d
7	y;ix;I	19	f;v
8	w;U;u	20	T;D
9	ow;o	21	k;g
10	aw	22	n
11	oy	23	p;b;m

Table 1: Phoneme to Viseme mapping

2.1. Data preparation

Video data from two female speakers, one Indian and one Croatian was recorded using a $Logitech\,HD\,pro\,webcam\,C920$. Video was recorded at $30\,fps$ and audio at $16\,kHz$ sampling rate at a resolution of $[1024 \times 576]$. Each speaker spoke twenty sentences from the CSRL English speech corpus [16], with 631 non unique phonemes, covering all the 48 phone classes in the phone set used. The audio was extracted and aligned with the transcription using HTK toolkit [17]. A second pass manual tuning was done on the automatically annotated data. Data for both the speakers was annotated using the same 48 class phone set, despite the perceivable difference between speakers for a particular phone, so as to assist in identification of phonetic distance.

The video was broken down into individual image frames for further analysis and each image frame tagged to a particular phone, based on the time-aligned transcription. Image frames for analysis were chosen in correspondence with the mid point of the respective phone segment. These static viseme image frames are visually intuitive, such as representing a bilabial phone when the lips are completely closed and for vowels when the lips are at maximum opening.

2.2. Computation of phonetic distance

Consider a speech utterance U, as a sequence of N phones p_i, such that, each phone belongs to one of the 48 classes.

$$\begin{aligned} U &\equiv (p_1, p_2, \ldots, p_n, \ldots, p_N) \quad \text{where} \\ p_n &\in \mathbb{P} \quad \text{where } \mathbb{P} \text{ is the set of 48 phones} \end{aligned} \tag{1}$$

We need to compute the distance between the utterances of the speakers, S_1 and S_2 at phone level. 13 MFCC features along with the first and and second derivative were extracted to form a feature vector of length 39 for each frame, with a window size of $20\,ms$ and overlap of $10\,ms$. Since the number of frames within a phonetic segment for the two speakers differed, the phonetic distance was computed as minimum cost alignment (MCA) between each segment using DTW algorithm [18]. The underlying assumption is that there is a one-to-one correspondence in the phone sequence for the same utterance for two different speakers. The MCA for the i_{th} phone of speaker S_1 and S_2 is computed as

$$PD_i = MCA(p_i^{S_1}, p_i^{S_2}) \tag{2}$$

Let the number of tokens present in each phone class be M, then the phonetic distance between speech in varying accents, for each phone class x is computed by aggregating the phonetic distance over the individual classes, as in Equation 3

$$PD_x = \frac{\sum_{j=1}^{M} PD_{x_j}}{M} \tag{3}$$

By using the phoneme-viseme mapping in Table 1, the phonetic distances for each viseme in the 24 viseme class, is computed as an average of phonetic distances computed for the phones mapped to that particular viseme class.

2.3. Computation of visemic distance

First step towards the computation of visemic distance is accurately identifying the *Region of Interest (ROI)*. We have restricted the ROI to size normalized lip/mouth region alone, for the purpose of our experiments. ROI was computed based on colour-based segmentation of the viseme images, followed by edge detection. A good contrast between the facial region and lip was ensured during recording, such that the mouth region was extracted during the color based segmentation. Image frames were segmented into three segments, using k-means clustering in $L*a*b*$ color space, where L is the luminance and a and b are the two color channels. Since, in the $L*a*b*$ model, the perceived color difference correspond to distances measured colorimetrically, segmentation of the mouth region which is a deeper shade of red, yields good segmentation results.

Edge detection is applied on the segment corresponding to red to narrow down the *Region of Interest (ROI)*, as shown in Figure 1.

The ROI now comprises the lips, mouth opening and sur-

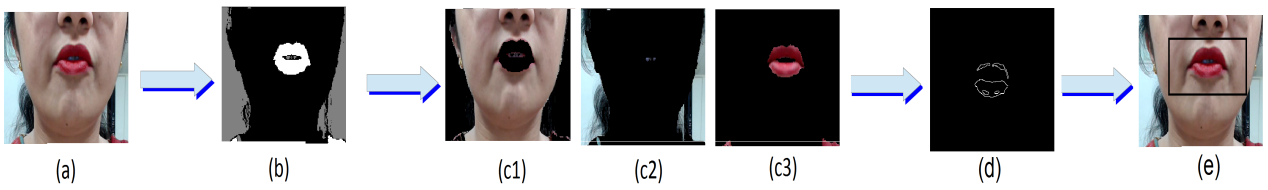

Figure 1: (a)Image frame (b)Segmented in L*a*b* color space (c)Three segments separated (d)Edge detected from c3, (e) ROI

rounding area and can be represented as

$$ROI \equiv \bigcup (Lips,\ Mouth\text{-}opening,\ Surrounding\text{-}region)$$
$$Non\text{-}lip\text{-}region \equiv \bigcup (Mouth\text{-}opening,\ Surrounding\text{-}region)$$
$$\implies ROI \equiv \bigcup (Lips,\ Non\text{-}lip\text{-}region)$$

Color based segmentation was in turn applied on the ROI, wherein two segments corresponding to lip region (LR) and non-lip region (NLR) were computed for each ROI image frame as shown in Figure 2

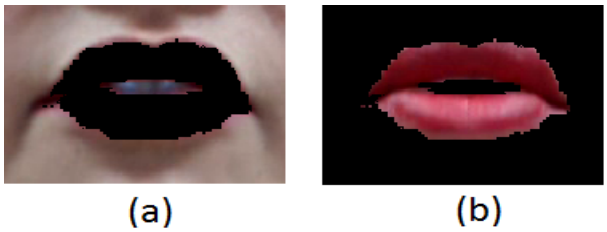

Figure 2: (a)Non-Lip region (b)lip region

Pixel density ρ, defined as number of pixels in the ROI for that particular segment, is computed for both LR and NLR. The ratio of ρ of the lip to non-lip region is used as metric VD for comparison of lip shapes of the two speakers for the same viseme.

$$VD_x = \frac{\rho_{LR}}{\rho_{NLR}} \quad \text{where } x \in \mathbb{V} \qquad (5)$$
$$\text{where } \mathbb{V} \text{ is the set of 24 visemes}$$

Using the above metric, an estimate of the mouth opening can be got for each viseme, since for a wider mouth opening (vowels), the ratio of ρ_{LR} to ρ_{NLR} would be higher than for visemes with closed lip shapes (silence).

The experimental setup and results are discussed in the next section.

3. Experimental results

Experimental setup and results for computation of phonetic distance PD, visual distance VD as discussed in the Section 2 and applicability to accent modification of visual subtitles are discussed.

Speech data was annotated using 48 phoneme classes. This 48 phoneme set was mapped to 24 viseme classes based on visual similarities as shown in Table 1. Phonetic distance for each of the 631 non unique phonemes was computed and aggregated over each phone in the 48 phoneme class and in turn aggregated into the 24 viseme classes. Maximum phonetic distance between an instance of a phone was 340.04, the maximum within-class standard deviation was 90.54

The metric VD, as in Equation 5, was computed for each viseme, for both Indian and Croatian speaker. The VD of the two speakers are not directly comparable considering parameters such as lip shape, size etc. Hence a relative grading for visemes within a speaker, based on VD was given. Difference in grade given for each viseme was computed to give the visual distance between visemes corresponding to the same speech sound. Table 2 shows the computation of the metric VD.

Viseme	VD (S1)	Rank (R1)	VD (S2)	Rank (R2)	Distance abs(R1-R2)
0	0.281	2	0.17	2	0
1	0.694	24	0.205	8	16
2	0.574	23	0.27	23	0
3	0.519	18	0.251	20	2
4	0.45	13	0.228	14	1
5	0.53	20	0.215	9	11
6	0.53	19	0.215	10	9
7	0.375	6	0.244	17	11
8	0.371	5	0.245	18	13
9	0.507	17	0.266	21	4
10	0.463	16	0.268	22	6
11	0.443	12	0.218	12	0
12	0.399	9	0.27	24	15
13	0.405	10	0.217	11	1
14	0.548	22	0.25	19	3
15	0.36	4	0.197	5	1
16	0.454	15	0.232	15	0
17	0.376	8	0.205	6	2
18	0.339	3	0.164	1	2
19	0.439	11	0.205	7	4
20	0.451	14	0.238	16	2
21	0.259	1	0.179	3	2
22	0.53	21	0.227	13	8
23	0.375	7	0.193	4	3

Table 2: Distance in visemic space

As can be seen from the Table 2, the visual distance for vowels such as /E/ and /I/ and consonants /T/ are on the higher side. The visual distance for silence, unvoiced consonants /k/, /g/ and bilabials /p/, /b/,/m/ are lower.

A comparison of PD and VD was done for each viseme in the similar fashion as explained above, by grading and comparing the rankings based on the distance in the phonetic and visual space. This similarity measure is an indication of how well the phonetic domain distances are matched with the visual domain. The similarity measure is expected to be closer to zero

if the distances in the visual domain are indeed caused by the difference in the auditory domain. Table 3 shows the distance computation between phonetic and visemic space for the top five farthest visemes, for Indian and Croatian speakers.

Viseme	Phoneme	Rank (VD)	Rank (PD)	Similarity measure
7	w;U;u	19	15	4
1	E	23	20	3
13	r	5	3	2
18	f;v	9	8	1
20	k;g	8	9	1

Table 3: Similarity measure between PD and VD

3.1. Application to visual subtitles

A comparison of the degree of similarity in the phonetic space and visual space paves way for a deeper understanding into the visual formation of speech sound for the same utterance in the same language(English) as spoken by natives of different countries. Our hypothesis is that the difference in visual appearance of a phoneme is based on the underlying phonetic characteristics. We design an experiment to be evaluated by a person who is trained to lip read Indian English. Word pairs with phonemes that were far apart in the phonetic domain are generated such that all other visemes except for the one being tested is kept constant, as shown in Table 4.

Target Phoneme	Word1(generated)	Word2
E	BATTLE /bEtəl/	MUDDLE /madəl/
U	SCHOOL /skUl/	SCOWL /skowl/
th	THROUGH /thru/	DREW /Dru/
T	TEASE /TIz/	THESE /dIz/

Table 4: Example of word pairs for evaluation of viseme formation

Viseme sequences for the word with the target phoneme is built using visemes from both Indian speaker and Croatian speaker. The evaluator is asked to chose the correct word from the pair, based on the viseme sequence The recognition word error rate is expected to be higher for the words generated using Croatian speaker's visemes. These lip movement sequences were evaluated by four subjects trained to lip read Indian English. Table 5 shows a comparative result.

4. Conclusions and future directions

We began our work with the understanding that a hearing impaired individual, trained to lip read Indian English is unable to lip read English spoken in other accents. The objective was to verify if the differences in the visual domain arise from the differences in the phonetic domain. The results in Section 3 strengthens our motivation to build a system that will be able to convert one accent of English to another in the visual domain. The above results can be applied to the tool that creates the visual subtitles, wherein using appropriate mapping between phonemes in a foreign accented English and visemes in Indian English, the hearing impaired user will be able to better lip read and understand the audio content. Further research directions include examining automatically viseme formations in larger audio-visual corpus such as the GRID corpus [19]. Towards this end, a parallel corpus needs to be built with Indian speakers. Impact of using dynamic visemes versus static visemes for analysis of correlation between visual and auditory domains [20] will be looked into. We would also like to explore the aspects such as visual representation of emotions, displaying distinction between voiced and unvoiced sounds for the same place of articulation.

Visual distance	Viseme	%Correct
Low	Croatian	61.1
High	Croatian	28.57
Low	Indian	77.7
High	Indian	71.4

Table 5: User evaluation of visemes for Indian English and Croatian English

5. Acknowledgements

The authors would like to acknowledge our colleagues for recording videos in Indian English and in Croatian English.

6. References

[1] C. Bhat, I. Ahmed, V. Saxena, and S. K. Kopparapu, "Visual subtitles for internet videos," in *InProc. SLPAT 2013 (Interspeech 2013 Satellite workshop on Speech and Language Processing for Assistive Technologies)*, Grenoble, France, 21-22 August, 2013, pp. 17–20.

[2] T. Paris, J. Kim, and C. Davis, "Visual speech speeds up auditory identification responses," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011, pp. 2469–2472.

[3] V. van Wassenhove, K. W. Grant, and D. Poeppel, "Visual speech speeds up the neural processing of auditory speech," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 4, pp. 1181–1186, 2005. [Online]. Available: http://www.pnas.org/content/102/4/1181.abstract

[4] E. Cvejic, J. Kim, and C. Davis, "Recognizing prosody across modalities, face areas and speakers: Examining perceivers sensitivity to variable realizations of visual prosody," *Cognition*, vol. 122, no. 3, pp. 442 – 453, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010027711002848

[5] J. Kim, E. Cvejic, and C. Davis, "Tracking eyebrows and head gestures associated with spoken prosody," *Speech Communication*, vol. 57, no. 0, pp. 317 – 330, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639313000691

[6] S. Taylor, B.-J. Theobald, and I. Matthews, "The effect of speaking rate on audio and visual speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 3037–3041.

[7] D. Burnham, V. Attina, and B. Kasisopa, "Auditory-visual discrimination and identification of lexical tone within and across tone languages," in *Auditory-Visual Speech Processing, AVSP 2011, Volterra, Italy, September 1-2, 2011*, 2011, pp. 37–42.

[8] A. Metallinou, C. Busso, S. Lee, and S. Narayanan, "Visual emotion recognition using compact facial representations and viseme information," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 2474–2477.

[9] S. Zhang, J. Jia, Y. Xu, and L. Cai, "Emotional talking agent: System and evaluation," in *Natural Computation (ICNC), 2010 Sixth International Conference on*, vol. 7, Aug 2010, pp. 3573–3577.

[10] C. Georgakis, S. Petridis, and M. Pantic, "Visual-only discrimination between native and non-native speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 4828–4832.

[11] P. van der Zande, A. Jesse, and A. Cutler, "Hearing words helps seeing words: A cross-modal word repetition effect," *Speech Communication*, vol. 59, no. 0, pp. 31 – 43, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639314000028

[12] D. Erdener and D. Burnham, "The relationship between auditory visual speech perception and language-specific speech perception at the onset of reading instruction in english-speaking children," *Journal of Experimental Child Psychology*, vol. 116, no. 2, pp. 120 – 138, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022096513000593

[13] L. E. Bernstein, E. T. J. Auer, J. K. Moore, C. W. Ponton, M. Don, and M. Singh, "Visual speech perception without primary auditory cortex activation," *NeuroReport*, vol. 13, pp. 311–315, 2002.

[14] C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar, "The natural statistics of audiovisual speech." *PLoS Computational Biology*, vol. 5, no. 7, 2009.

[15] L. Cappelletta and N. Harte, "Phoneme-to-viseme mapping for visual speech recognition." in *ICPRAM (2)*, 2012, pp. 322–329.

[16] S. K. Kopparapu, https://sites.google.com/site/awazyp/data/speechcorpus, viewed March 2015.

[17] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge University Press, 2006.

[18] J. Vavrek, M. Pleva, and J. Juhar, "Tuke mediaeval 2012: Spoken web search using dtw and unsupervised svm." in *MediaEval*, ser. CEUR Workshop Proceedings, M. A. Larson, S. Schmiedeke, P. Kelm, A. Rae, V. Mezaris, T. Piatrik, M. Soleymani, F. Metze, and G. J. F. Jones, Eds., vol. 927. CEUR-WS.org, 2012.

[19] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[20] S. L. Taylor, M. Mahler, B.-J. Theobald, and I. Matthews, "Dynamic units of visual speech," in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, ser. SCA '12. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2012, pp. 275–284. [Online]. Available: http://dl.acm.org/citation.cfm?id=2422356.2422395