THE EFFECT OF SPEAKING RATE ON AUDIO AND VISUAL SPEECH

Sarah Taylor [1] *Barry-John Theobald* [2] *Iain Matthews* [1]

[1] Disney Research, Pittsburgh, PA
[2] University of East Anglia, Norwich, UK

ABSTRACT

The speed that an utterance is spoken affects both the duration of the speech and the position of the articulators. Consequently, the sounds that are produced are modified, as are the position and appearance of the lips, teeth, tongue and other visible articulators. We describe an experiment designed to measure the effect of variable speaking rate on audio and visual speech by comparing sequences of phonemes and dynamic visemes appearing in the same sentences spoken at different speeds. We find that both audio and visual speech production are affected by varying the rate of speech, however, the effect is significantly more prominent in visual speech.

Index Terms— Audio-visual speech, speaking rate, dynamic visemes

1. INTRODUCTION

Speech production is a complex function involving a large number of interacting processes. We present an investigation into the influence of speaking rate on the way that a person articulates an utterance. We consider the affect of speaking rate on both the acoustic and the visual aspects of speech. This is an important consideration for both accurate speech recognition (audio, visual and audio-visual) and natural speech synthesis (audio and visual), yet the influence of speaking rate is often overlooked.

Speaking rate can be defined as the number of words or syllables spoken over a unit of time. Speaking rate depends on many characteristics of the speaker, including their age, gender, physiology and psychological state. For example, on average, younger people talk faster than older people [1], males talk faster than females [1], and people speak faster when they are angry than when they are sad [2]. People also tend to talk at a faster rate when speaking to a person who is familiar to them and speaking rate increases as the length of the intended utterance increases [1]. Variations in speaking rate frequently occur in natural speech.

Speech spoken at different rates influences the duration of the segments and also the acoustic properties of the utterance. For example, as speaking rate increases, the pitch range decreases [3] and it is typical for articulators to undershoot targets to the extent that all vowels are reduced to a schwa [4]. This may in part be due to physiological constraints, since muscles need a minimum amount of time to contract. Speech spoken at different rates is not the same as "typical" speech spoken more quickly or more slowly.

Little is known about the effect that varying speaking rate has on the movement of the visible articulators, so called *visual speech*. It is clear that fast speech is not simply slow speech sped up. Instead, sequential segments are merged, boundaries are blurred and some segments are deleted or inserted. The work described in this paper measures the extent to which visual and acoustic information is affected by variable speaking rates by comparing sequences of phonemic and visemic labels corresponding to repetitions of sentences spoken fast, normal and slowly.

2. PHONEMES AND VISEMES

Phonemes are well defined linguistic units of acoustic speech. They represent the contrastive sounds of a language and so can be used to unambiguously transcribe speech utterances. However, the visual equivalent of the phonemes is not so well defined. For years, *visemes* ("visual phonemes") were proposed as the units of visual speech [5], identified by clustering phonemes based on their visible articulator configuration such that phonemes produced with a similar pose were grouped to form a single viseme class. This clustering has been performed both subjectively [5–11] and objectively [12–16] using a range of different speakers, stimuli, and recognition/classification tasks. However, no unequivocal mapping from phonemes to (phoneme cluster) visemes exists. This is because there is no simple many-to-one mapping from phonemes to visual speech. Static visemes do not account for visual coarticulation, which is the influence of neighboring speech on the position of the articulators. Coarticulation causes the lip pose for the same sound to appear very different visually depending on the context in which it is embedded (see Figure 1) and at times the articulation of some sounds may not be visible at all. For this reason, the traditional definition of a viseme functions as a poor model for a unit of visual speech.

A more realistic model of visual speech is *dynamic visemes* [17]. Dynamic visemes are speech *movements* rather than static poses and they are learnt by clustering

Fig. 1: A selection of movie frames during the articulation of /t/ illustrating the variability of articulator poses due to coarticulation.

visual speech independently of the underlying phoneme labels. Given some training video containing speech, the visible articulators are tracked and parameterized into a low-dimensional space. This parameterization is then automatically segmented by identifying salient points to give a series of short, non-overlapping gestures. These salient points are visually intuitive and fall at locations where the articulators change direction, for example as the lips close during a bilabial, or the peak of the lip opening during a vowel. The speech gestures identified by this segmentation are then clustered to form dynamic viseme groups, such that *movements* that look very similar appear in the same class. Identifying visual speech units in this way is beneficial as the set of dynamic visemes describes all of the ways in which the visible articulators move during speech. For the remainder of this paper *visemes* refers to *dynamic visemes* as defined in [17].

3. DATA CAPTURE AND METHODS

3.1. Audio-Visual Speech Database

A dataset was captured containing an actor speaking 10 sentences from the TIMIT sentence list [18] at 3 speeds (slow, normal and fast), each repeated 10 times ($10 \times 3 \times 10 = 300$ utterances). The prompts were presented in a randomised order in which the sentences and speaking rates were varied. The speaking rates of the uttered sentences were checked and were in accordance with the speeds that the actor was asked to speak. The video was recorded at 29.97 frames per second at a resolution of 1920 by 1080 progressive scan using a Sony PMW-EX3 camera. Audio was synchronously captured at a sampling rate of 48kHz.

3.2. Data Preparation

To generate dynamic viseme labels the jaw and lips were first tracked and parameterized for each video frame using an active appearance model (AAM) [19]. Speech gestures were identified by automatically segmenting in AAM space based on zero crossings from negative to positive in the derivative of the gradient magnitude. The clustering was performed in super-feature space using a graph-based clustering method [20], generating 102 dynamic viseme classes. The number of classes was determined as the point where the mean squared difference between the super-features within a cluster and the respective cluster median does not change significantly after increasing the number of clusters. See [17] for more details.

As a result of the clustering process, each of the 300 sentences were labelled in terms of their dynamic viseme class labels (1-102) and the sentences were phonetically labelled using ARPABet notation manually.

	Slow	Normal	Fast
Slow	89.2 ± 2.9	87.5 ± 3.6	84.8 ± 3.5
Normal	86.7 ± 4.0	90.6 ± 3.0	88.8 ± 3.1
Fast	83.2 ± 4.5	88.4 ± 3.7	88.5 ± 3.0

Table 1: The means and standard deviations of phonemic similarity (as a percentage) measured across speech sequences spoken at different rates.

3.3. Measuring the Effect of Variable Speaking Rate

The effect of speaking rate on acoustic speech is measured by calculating the similarity between phoneme sequences for repetitions of the same sentences spoken at different speeds. The alignment of the phonemic transcriptions was achieved using forced alignment in the hidden Markov model toolkit (HTK) [21], and the phonemic similarity was calculated using the inverse Levenshtein distance [22]:

$$\text{Similarity} = \frac{N - D - S - I}{N} \times 100\%, \qquad (1)$$

where N is the total number of labels in the reference sentence, and D, S and I are the number of deletions, substitutions and insertions respectively.

To measure the effect of speaking rate on visual speech, for each sentence the viseme labels were aligned with all other repetitions of the sentence using dynamic programming, and the visemic similarity was measured using Equation 1. In both the audio and the visual cases the similarity was averaged across the 10 sentences and 10 repetitions for each of the three speaking rates.

4. RESULTS

The mean and standard deviations of phonemic and visemic similarities for sentences spoken at different speaking rates are shown in Tables 1 and 2 respectively. The intersection of a row/column shows the percentage similarity for the corresponding speaking rates. Unsurprisingly, the phonemic similarities for acoustic speech are high as the same phones tend to be used to produce different repetitions of the same utterance regardless of speaking rate. However, the results in Table 1 indicate that even when speaking the same sentence at the same speaking rate, the phones uttered are not identical.

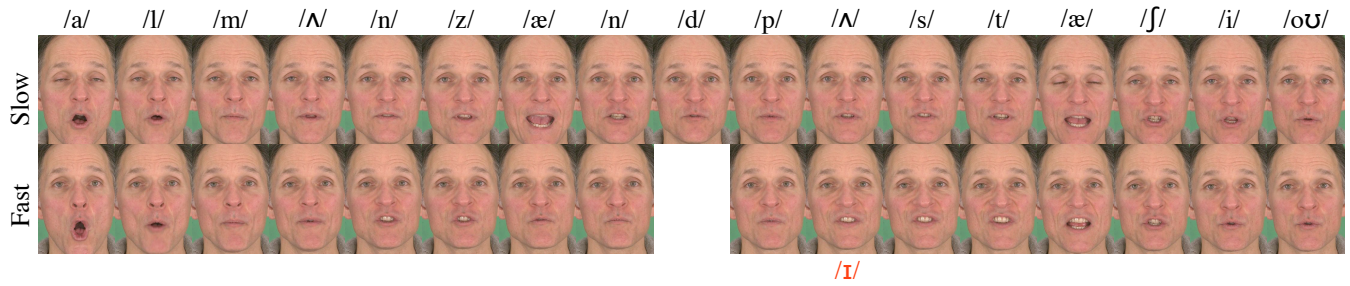


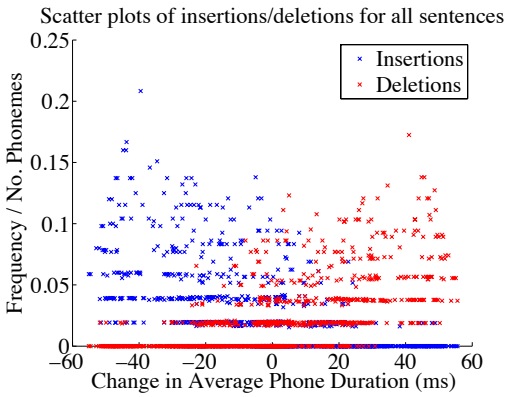
Fig. 2: Frames taken from the mid-point of each phone segment for the phrase "Almonds and pistachio [nuts]". The top row is from a slow repetition, and the bottom row is from a fast repetition. No image is shown where a deletion occurs and a phoneme substitution is shown in red.

	Slow	Normal	Fast
Slow	53.9 ± 7.9	45.5 ± 7.6	27.4 ± 6.7
Normal	32.7 ± 13.3	57.6 ± 8.4	39.8 ± 9.6
Fast	-11.7 ± 19.6	23.8 ± 17.4	57.6 ± 8.7

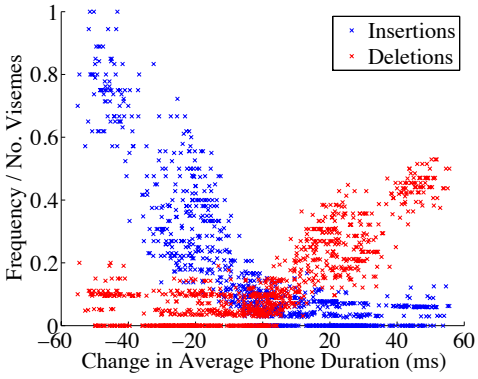
Table 2: The means and standard deviations of visemic similarity (as a percentage) measured across sentences spoken at different rates. Note the negative value comparing slow and fast speech due to a large number of speech units present in slow speech which are missing from fast speech.

The lowest similarity is measured between sentences that are spoken at a fast rate and those that are spoken slowly. For this speaker, the phones that are most likely to be dropped from speech that is spoken at a higher rate are /h, t, j/, where deletions occur 40%, 20%, and 18% of the time respectively. The most common consonant substitutions for faster speech are /z/→/s/, /t/→/d/ and /θ/→/ð/, occurring 16%, 9% and 5% of the time respectively. Vowels generally become less well defined and /u/, /ʊ/ and /æ/ often reduce to /ʌ/ (12%, 12% and 9% of the time respectively), and /ɔ/ is substituted for /ɑ/ 11% of the time. A selection of aligned phoneme sequences for the sentence "Almonds and pistachio nuts are not so high in oil, but are rich in protein" is shown in Table 3.

For visual speech the visemic similarities are much lower, indicating that different lip motions are used to produce the same words when speaking at different rates. In Table 2 note the negative value comparing slow and fast visual speech, which suggests that a large number of dynamic visemes present in slow speech are missing from fast speech. This is confirmed in Table 4, which shows a selection of viseme sequences for repetitions of a sentence which have been aligned for visualization. In all cases, the viseme sequences for sentences spoken at a particular speaking rate are more similar to others spoken at the same speed than those that are spoken at different speeds, and the faster that a sentence is produced, the fewer visemes are used. Furthermore, the results indicate that visual speech is influenced more by the effect of variable speaking rate than acoustic speech as the difference



(a) Audio



(b) Video

Fig. 3: Normalized frequency of insertions and deletions for each sentence compared with all of its repetitions. Each sample represents one comparison. Since each sentence is compared with 10 repetitions for each of the 3 speeds, there are 300 samples in each colour. The x-axis represents the difference between the mean phoneme duration of a sequence and another repetition to which it is aligned (see Equation 2). The number of insertions (red) and deletions (blue) are normalized for sequence length.

	Slow
1	ɑ l m ʌ n z - æ n p ɪ s t æ ʃ i oʊ n ɑ t s - ɑ r n ɑ t s oʊ h aɪ ɪ n ɔɪ l - b ʌ t ɑ r ɪ tʃ - ɪ n p r oʊ t i n
2	ɑ l m ʌ n z - æ n d p ɪ s t æ ʃ i oʊ n ʌ t s - ɑ n ɑ t s oʊ h aɪ ɪ n ɔɪ l - b ʌ t ɑ r ɪ tʃ - ɪ n p r oʊ t i n
	Normal
1	ɑ l m ʌ n d z ʌ n p ɪ s t æ ʃ j ʊ n ɑ t s ɑ r n ɑ t s oʊ h aɪ ɪ n ɔɪ l b ʌ d ɑ r ɪ tʃ ɪ n p r oʊ t i n
2	ɑ l m ʌ n z ʌ n p ɪ s t æ ʃ j ʌ n ɑ t s ɑ r n ɑ t s oʊ h aɪ ɪ n ɔɪ l b ʌ d ɑ r ɪ tʃ ɪ n p r oʊ t i n
	Fast
1	ɑ l m ʌ n z ʌ n p ɪ s t æ ʃ oʊ n ʌ t s ɑ n ɑ t s oʊ h aɪ ɪ n ɔɪ l b ʌ t ɑ r ɪ tʃ ɪ n p r oʊ t i n
2	ɑ l m ʌ n z n p ʌ s t æ ʃ oʊ n ʌ t s ɑ r n ɑ t s oʊ h aɪ ɪ n ɔɪ l b ʌ d ɑ r ɪ tʃ ɪ n p r oʊ t i n

Table 3: A selection of aligned phoneme sequences for the sentence "Almonds and pistachio nuts are not so high in oil, but are rich in protein" spoken at different rates. A dash (-) denotes a short pause.

	Slow
1	5 56 67 11 83 56 62 54 93 97 92 43 99 27 64 50 55 13 49 58 98 53 88 19 99 74 6 89 93 35 45 86 79 41
2	5 56 67 11 83 56 62 54 93 97 66 37 84 29 74 64 72 101 55 13 49 63 58 98 44 19 27 6 80 22 84 18 35 45 86 71 73
	Normal
1	5 56 61 56 62 54 100 80 66 67 84 87 91 55 13 49 58 98 84 88 19 38 80 93 35 45 86 71 52
2	5 56 61 56 62 54 100 80 87 67 84 87 1 13 49 58 98 44 88 38 89 93 35 101 86 71 52
	Fast
1	5 75 8 62 83 80 87 84 87 1 83 80 70 88 6 89 93 35 81 71 52
2	5 96 61 8 62 83 80 87 84 87 1 83 58 70 88 100 89 61 35 81 79 41

Table 4: A selection of aligned viseme sequences for the sentence "Almonds and pistachio nuts are not so high in oil, but are rich in protein" spoken at different rates.

in visemic similarity is larger and the number of units in a sequence is more variable.

Figure 2 shows movie frames taken from the mid-point of each phone segment for the phrase "Almonds and pistachio nuts" for slow speech (top row) and fast speech (bottom row). A blank image appears where a phone is not articulated. Note the difference in lip pose across the two speeds, especially in the build up to /p/. During the fast repetition the lips move towards lip closure much earlier than during the slow repetition.

Figure 3 shows for phone (top) and viseme (bottom) sequences the number of insertions and deletions as a function of the change in average phone duration, which is calculated using:

$$\Delta\bar{d}^{ab} = \frac{1}{N^a}\sum_{i=1}^{N^a} d_i^a - \frac{1}{N^b}\sum_{j=1}^{N^b} d_j^b \,, \qquad (2)$$

where N is the number of phonemes in a sentence and d_i^a and d_j^b are the durations of phonemes i and j in repetitions a and b respectively. Each data point represents a comparison between two repetitions of a sentence and shows the number of insertions (blue) and deletions (red) of the aligned sequences normalized for sequence length. In both graphs the number of insertions increases when aligning slow speech to fast ($\Delta\bar{d}^{ab} < 0$) and the number of deletions increases when aligning fast speech to slow ($\Delta\bar{d}^{ab} > 0$). However, the trend is much more prominent for visual speech, where the number of data points in the top left of the figure can double when the speech is spoken slowly.

5. DISCUSSION AND FUTURE WORK

Whilst differences in visual speaking style are to be expected as speaking rate varies, the magnitude of the differences in Table 2 are surprising. Significantly, this confirms that speaking fast is not the same as slow speech spoken more quickly. Coarticulation affects speech production for different speaking rates since the visual gestures used to produce the same sentence at different speeds is very different and the differences are much more prominent visually than acoustically.

In this work, dynamic visemes are learned for a single speaker and are the best units for describing the speech movements specific to this speaker. However, phonemes are generic and can be conferred across speakers, enabling a clearer evaluation of the variability due to speaking rate. A future goal is to learn and publish a full, speaker-independent set of dynamic visemes.

We are in the process of extending this study to consider a larger corpus of speech, including more speakers, and we will also investigate the influence of emotion on speech production. Our longer term goal is to better understand visual speech articulation such that speech animation can be adapted automatically so that the visual appearance looks correct given the intended speaking rate and style.

6. REFERENCES

[1] Jiahong Yuan, Mark Liberman, and Christopher Cieri, "Towards an integrated understanding of speaking rate in conversation.," in *Interspeech*, 2006, pp. 541–544.

[2] K. Sreenivasa Rao and Shashidhar G. Koolagudi, "Robust emotion recognition using speaking rate features," in *Robust Emotion Recognition using Spectral and Prosodic Features*, SpringerBriefs in Electrical and Computer Engineering, pp. 85–94. Springer New York, 2013.

[3] Aijun Li and Yiqing Zu, "Speaking rate effects on discourse prosody in standard chinese," *Speech Prosody*, pp. 449–452, 2008.

[4] B. Lindblom, "Spectrographic study of vowel reduction," *The Journal of the Acoustical Society of America*, vol. 35, no. 11, pp. 1773–1781, 1963.

[5] C. Fisher, "Confusions among visually perceived consonants," *Journal of Speech and Hearing Research (JSHR)*, vol. 11, pp. 796–804, 1968.

[6] E. T. Auer and L. E. Bernstein, "Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness," *Journal of the Acoustical Society of America (JASA)*, vol. 102, pp. 3704–3710, Dec. 1997.

[7] S.A. Lesner, S.A. Sandridge, and P.B. Kricos, "Training influences on visual consonant and sentence recognition," *Ear and Hearing*, vol. 8, no. 5, 1987.

[8] C. A. Binnie, P. L. Jackson, and A. A Montgomery, "Visual intelligibility of consonants: A lipreading screening test with implications for aural rehabilitation," *Journal of Speech and Hearing Disorders*, vol. 41, pp. 530–539, 1976.

[9] B. Lidestam and J Beskow, "Visual phonemic ambiguity and speechreading," *Journal of Speech, Language and Hearing Research (JSLHR)*, vol. 49, pp. 835–847, August 2006.

[10] S.A. Lesner and P.B. Kricos, "Visual vowel and diphthong perception across speakers," *Journal of the Academy of Rehabilitative Audiology (JARA)*, pp. 252–258, 1981.

[11] B.E. Walden, R.A. Prosek, A.A. Montgomery, C.K. Scherr, and C.J. Jones, "Effects of training on the visual recognition of consonants," *Journal of Speech, Language and Hearing Research (JSLHR)*, vol. 20, no. 1, pp. 130–145, 1977.

[12] Alan J. Goldschen, Oscar N. Garcia, and Eric Petajan, "Continuous optical automatic speech recognition by lipreading," in *Proceedings of the 28th Asilomar Conference on Signals, Systems, and Computers*, 1994, pp. 572–577.

[13] Timothy J. Hazen, Kate Saenko, Chia-Hao La, and James R. Glass, "A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments," in *Proceedings of the International conference on Multimodal Interfaces (ICMI)*, New York, NY, USA, 2004, pp. 235–242, ACM.

[14] J. Melenchón, J. Simó, G. Cobo, and E. Martínez, "Objective viseme extraction and audiovisual uncertainty: Estimation limits between auditory and visual modes," in *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)*, 2007.

[15] N.M. Brooke and P.D. Templeton, "Classification of lip-shapes and their association with acoustic speech events," in *The ESCA Workshop on Speech Synthesis*, September 1990, pp. 245–248.

[16] José Mario De Martino, Léo Pini Magalhães, and Fábio Violaro, "Facial animation based on context-dependent visemes," *Journal of Computers and Graphics*, vol. 30, no. 6, pp. 971 – 980, 2006.

[17] Sarah Taylor, Moshe Mahler, Barry Theobald, and Iain Matthews, "Dynamic units of visual speech," in *ACM/ Eurographics Symposium on Computer Animation (SCA)*, 2012, pp. 275–284.

[18] NIST, "The DARPA TIMIT acoustic-phonetic continuous speech corpus (TIMIT)," [CD-ROM], November 1988.

[19] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 23, no. 6, pp. 681–685, June 2001.

[20] G. Karypis, *CLUTO — A clustering toolkit*, University of Minnesota, Department of Computer Science, Minneapolis, April 2002.

[21] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*, 2006.

[22] V.I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics — Doklady, Cybernetics and Control Theory*, vol. 10, no. 8, pp. 707–710, 1966.