A MOUTH FULL OF WORDS: VISUALLY CONSISTENT ACOUSTIC REDUBBING

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

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

ABSTRACT

This paper introduces a method for automatic *redubbing* of video that exploits the many-to-many mapping of phoneme sequences to lip movements modelled as *dynamic visemes* [1]. For a given utterance, the corresponding dynamic viseme sequence is sampled to construct a graph of possible phoneme sequences that synchronize with the video. When composed with a pronunciation dictionary and language model, this produces a vast number of word sequences that are in sync with the original video, literally putting plausible words into the mouth of the speaker. We demonstrate that traditional, one-to-many, static visemes lack flexibility for this application as they produce significantly fewer word sequences. This work explores the natural ambiguity in visual speech and offers insight for automatic speech recognition and the importance of language modeling.

Index Terms— Audio-visual speech, dynamic visemes, acoustic redubbing.

1. INTRODUCTION

Redubbing is the process of replacing the audio track in a video. This paper focuses on redubbing speech, which involves substituting an utterance with another that, when composited with the original video, appears to be in sync with the movements of the visible articulators. The primary application of speech redubbing is translating movies, television shows and video games for audiences that speak a different language to the original recording. It is also common to replace speech with different dialogue from the same language. For example, a movie may be edited for television by redubbing offensive phrases. A more restricted approach aims to generate speech that is perfectly in sync and *consistent* with the visual speech. Typically, the new dialogue is meticulously scripted in an attempt to select words that approximate the lip-shapes in the video, and it requires skill on the part of the voice actor to ensure their new voice recording actually synchronizes well with the existing movements of the lips, so this is a challenging task.

Automatic speech redubbing is an unexplored area of research. It shares similarities to automatic recognition of visual speech reading in that it involves decoding word sequences from a visual speech signal. However, the goal of this work is to suggest visually consistent *alternative* word sequences rather than predict the original speech. This paper proposes a novel method for automatic speech redubbing using dynamic visemes to represent the relationship between visible articulator motion and the underlying acoustic units. Dynamic visemes capture distributions of phonemes, so are a more accurate and richer source of information than the traditional, static visemes. A phoneme graph is constructed from the dynamic viseme sequence of an utterance, which is searched for word sequences. The word sequences are ranked using a language model. We compare this approach to using traditional, static visemes for redubbing.

2. REPRESENTING VISUAL SPEECH

2.1. Visemes

Until recently, *visemes* ("visual phonemes") were proposed as the units of visual speech [2]. They were identified by grouping phonemes based on their visual similarity such that phonemes that are produced with a similar visible articulator configuration formed a single viseme class. Typical viseme groupings include the closed mouth class, /p, b, m/, and the lower lip tuck class, /f, v/. See Table 1 for some example viseme mappings. Viseme classes are formed either subjectively [2–9] or objectively [10–14] using a range of different speakers, stimuli, and recognition/classification tasks. However, no unequivocal mapping from phonemes to visemes exists in terms of both the number and composition of the classes. This is because there is no simple many-to-one mapping from phonemes to visual speech. Visemes defined as phoneme clusters 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 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 unit of visual speech.

2.2. Dynamic Visemes

A better model of visual speech is *dynamic visemes* [1]. Dynamic visemes are speech *movements* rather than static poses and they are derived from visual speech independently of the underlying phoneme labels. Given a video containing a talking face, dynamic visemes are learned as follows: 1) Track the visible articulators and parameterize into a low-dimensional space. 2) Automatically segment the parameterization by identifying salient points to give a series of short, non-overlapping gestures. The salient points identified in this step 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. 3) Cluster the speech gestures identified by Step 2 to form dynamic viseme groups, such that *movements* that look very similar appear in the same class. More details can be found in [1]. Identifying visual speech units in this way is beneficial as the set of dynamic visemes describes all of the distinct ways in which the visible articulators move during speech. Additionally, dynamic visemes are learned entirely from visual data and no assumptions are made regarding the relationship to the acoustic phonemes.

Taylor et al. [1] found that dynamic visemes are a good unit of visual speech for generating realistic speech animation. Furthermore, they showed that the same sentences spoken at different speeds are transcribed into essentially the same (traditional) visemes even though the underlying lip shapes used to produce the speech are very different [15]. This paper uses dynamic visemes to look at the problem of redubbing. For the remainder of this paper *static visemes* refers to traditional units formed by clustering phonemes (Section 2.1) and *dynamic visemes* refers to those described in this section as defined by Taylor et al. [1].

2.3. The Many-to-Many Relationship between Phoneme Sequences and Dynamic Visemes

There is a complex many-to-many mapping between phoneme sequences and dynamic visemes. Different gestures that correspond to the same phoneme sequence can be clustered into multiple classes since they can appear distinctive when spoken at variable speaking rates or in different contexts. Conversely, a dynamic viseme class contains gestures that map to different phoneme strings. A valuable property of dynamic visemes is that they provide a probabilistic mapping from speech movements to phoneme sequences (and vice versa) by evaluating the probability mass distributions. Fig. 1 shows sample distributions for three dynamic viseme classes. The work described in this paper takes advantage of this property by sampling the phoneme string distributions for a known sequence of dynamic visemes.

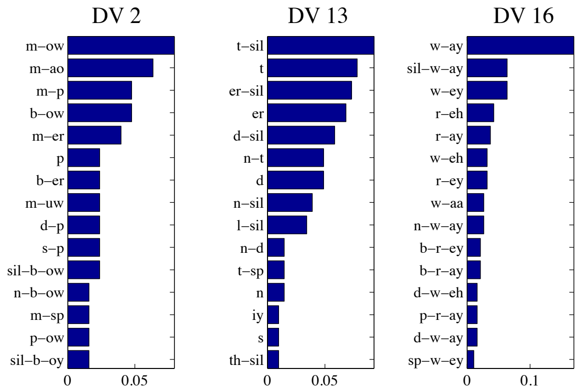


Fig. 1: A sampling of the phoneme string distributions for three dynamic viseme classes.

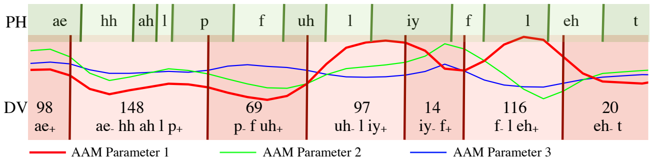


Fig. 2: Phonemes (PH) and dynamic visemes (DV) for the phrase "a helpful leaflet". Note that PH and DV boundaries need not align, so phonemes that are intersected by the DV boundaries are assigned a context label.

2.4. Phonetic Boundary Context

A dynamic viseme class represents a cluster of similar visual speech gestures, each corresponding to a phoneme sequence in the training data. Since these gestures are derived independently of the phoneme segmentation, the visual and acoustic boundaries need not align due to the natural asynchrony between speech sounds and the corresponding facial movements. Taylor et al. [1] found that 90% of the visual speech gestures spanned between two and seven phones or partial phones. For better modeling in situations where the boundaries are not aligned, the boundary phonemes are annotated with contextual labels that signify whether the gesture spans the beginning of the phone (p_+), the middle of the phone (p_*) or the end of the phone (p_-) (see Fig. 2).

3. DATA CAPTURE AND PREPARATION

3.1. Audio-Visual Speech Database

In this work dynamic visemes were learned from the KB-2K dataset, which contains an actor reciting 2542 phonetically balanced sentences from the TIMIT sentences. The video was captured in full-frontal view at 29.97 fps at 1080p using a Sony PMW-EX3 camera over 8 hours. The actor was asked to talk in a neutral speaking style and maintain, as far as possible, a fixed head pose. All sentences were annotated manually using the phonetic labels defined in the Arpabet phonetic transcription code. The jaw and lips were tracked and parameterized using active appearance models (AAMs) providing a compact 20D feature vector describing the variation in both shape and appearance at each video frame. These were automatically segmented into ≈50000 visual speech gestures and clustered to form 150 dynamic viseme classes using the approach in [1].

4. METHODS

The goal is to generate a set of visually consistent phoneme sequences with corresponding durations that, when played back with the original video of a person speaking, appear to synchronize with the visible articulator motion.

4.1. Dynamic Visemes to Phonemes

Given the dynamic viseme sequence, $\mathbf{v} = v_1, \ldots, v_n$, the goal is to produce a set of word sequences, W, where $W_k = w_{(k,1)}, \ldots, w_{(k,m)}$. The first step involves constructing a (directed acyclic) graph which models all valid phoneme paths through the dynamic viseme sequence. A graph node is added for every unique phoneme sequence in each dynamic viseme in the sequence. Edges are then positioned between nodes of consecutive dynamic visemes where a transition is valid, constrained by the contextual labels assigned to the boundary phonemes as described in Section 2.4. For example, if contextual labels suggest that the beginning of a phoneme appears at the end of one dynamic viseme, the next should contain the middle or end of the same phoneme, and if the entire phoneme appears, the next gesture should begin from the start of a phoneme. The probability of the phoneme string with respect to its dynamic viseme class is also stored in each node.

4.2. Phonemes to Words

The next step is to search the phoneme graph for sequences that form complete strings of words. For efficient phoneme sequence-to-word lookup a tree-based index is constructed offline, which allows any phoneme string, $\mathbf{p} = p_1, \ldots, p_j$, as a search term and returns all matching words. This is created using the CMU Pronouncing Dictionary [16].

A left-to-right breadth first search algorithm is used to evaluate the phoneme graph. At each node, all word sequences that correspond to all phoneme strings up to that node are obtained by exhaustively and recursively querying the dictionary with phoneme sequences of increasing length up to a specified maximum. The probability of a word sequence is calculated using:

$$P(\mathbf{w} \mid \mathbf{v}) = \sum_{i=1}^{m} \log P(w_i \mid w_{i-1}) + \sum_{j=1}^{n} \log P(\mathbf{p} \mid v_j). \quad (1)$$

$P(\mathbf{p} \mid v)$ is the probability of phoneme sequence $\mathbf{p}$ with respect to the viseme class and $P(w_i \mid w_{i-1})$ is calculated using a word bigram language model trained on the Open American National Corpus [17]. To account for data sparsity, the probabilities are smoothed using Jelinek-Mercer interpolation [18]:

$$P(w_i \mid w_{i-1}) = \lambda \frac{C(w_i w_{i-1})}{C(w_i) + V} + (1 - \lambda) \frac{C(w_i) + 1}{N} \quad (2)$$

where N is the number of words in the dataset and V is the size of the vocabulary.

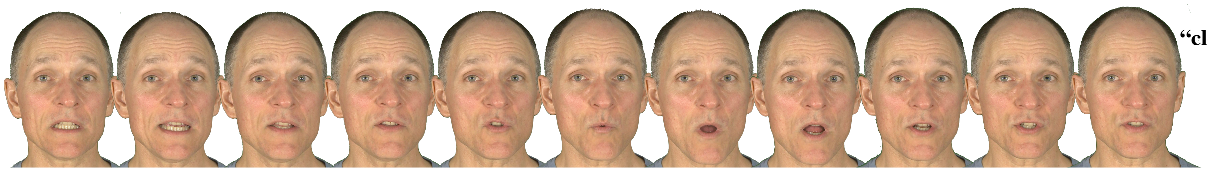
A breadth first graph traversal allows for Equation 1 to be computed for every viseme in the sequence allowing optional thresholding to prune low scoring nodes and increase efficiency. The algorithm also allows for partial words to appear at the end of a word sequence when evaluating mid-sentence nodes. The probability of a partial word is the maximum probability of all words that begins with the phoneme substring [19], $P(w^p) = \max_{w \in \mathbf{w}^p} P(w)$, where $\mathbf{w}^p$ is the set of words that start with the phoneme sequence w^p, $\mathbf{w}^p = \{w \mid w_{(1 \ldots k)} = w^p\}$. If all paths to a node cannot comprise a word sequence, it is removed from the graph. A complete word sequence is required when the final nodes are evaluated.

5. ALTERNATIVE DIALOGUE FROM STATIC VISEMES

For comparison, the use of traditional, many-to-one *static* visemes for redubbing was also explored. Each phoneme in a sequence is substituted with another from the same static viseme class to generate alternative word sequences. Unfortunately, most phoneme-to-static viseme mappings are incomplete and only consider a subset of the phonemes, typically clustering only consonants under the assumption that vowels form their own class. Table 1 shows two (mostly) complete but very different mappings as defined by Jeffers and Barley (JB) [3] and Parke and Waters (PW) [20] who identified 11 and 18 viseme classes respectively. On average, mapping JB contains 3.5 phonemes per class, and PW just 1.9. For an average 10 phoneme phrase, JB allows for $\approx 2.8 \times 10^5$ phoneme permutations in which to search for word sequences, whereas PW allows for only 357 permutations, providing a more constrained search space. The efficacy of this approach therefore depends highly on the phoneme-to-viseme mapping used.

6. RESULTS

A set of short phrases ranging in length from 5 to 15 phonemes were identified in the KB-2k dataset for which the dynamic viseme sequence is known, and 50 of each length were sampled. The number of unique word sequences generated using the methods described in Section 4 with dynamic visemes (DV), Jeffers and Barley's static visemes (JB) and Parke and Waters' static visemes (PW) was calculated. The phoneme-to-word search was performed with no graph pruning such



“clean swatches”
“likes swats”
“then swine”
“need no pots”
“tikes rush”

Fig. 3: Video frames from the phrase “clean swatches” and a sample of visually consistent alternative phrases identified by the dynamic viseme-to-word sequence algorithm.

Author	Static Viseme Mapping				
Parke and Waters [20]	/p b m/	/f v/	/dh th/	/ch jh sh zh/	
	/s z/	/d n t/	/l/	/g k ng hh er/	
	/r/	/w uh/	/uw/	/ao/	/iy
	/aa/	/ae/	/ah/	/ih y/	/eh/
Jeffers and Barley [3]	/p b m/	/f v/	/dh th/	/ch jh sh zh/	
	/s z/	/d n t l/	/g k ng/	/aw/	/oy ao/
	/er ow r w uh uw/		/aa ae ah ay eh ey ih iy y/		

Table 1: Many-to-one phoneme-to-static viseme mappings defined by Parke and Waters [20] and Jeffers and Barley [3].

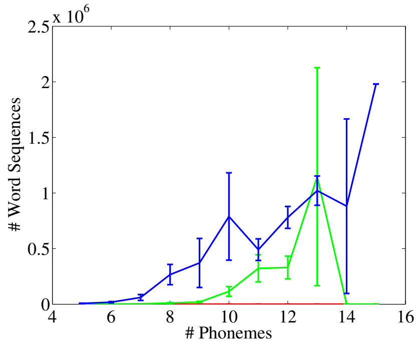


Fig. 4: The average number of word sequences retrieved for variable length phrases for Parke visemes (red), Jeffers visemes (green) and dynamic visemes (blue).

that a comprehensive list of word sequences could be attained for each method. Fig. 4 shows the average number of word sequences retrieved for variable length phrases using each method. It is clear that DV can produce a larger number of alternative word sequences than both static viseme approaches.

When using Parke and Waters’ visemes, 58% of the time the algorithm fails to find any valid word sequences because the search space is too small. Fail cases also occur using Jeffers and Barley’s mapping 17% of the time. Both mappings contain classes which are a mixture of vowels and consonants, increasing the likelihood of producing a linguistically invalid phoneme sequence since replacing a vowel with a consonant can produce a long string of consecutive consonants, which is uncommon in the English language. This is less likely to occur using dynamic visemes as naturally occurring phoneme *sequences* are contained within the units, and boundary context labels enforce valid transitions between units. Additionally, the static viseme approach is limited to word sequences with the same number of phonemes as the original speech.

To gauge how well a word sequence synchronizes with lip motion in the video, the Festival Speech Synthesis System [21] is used to generate a new audio track containing the phoneme string corresponding to the word sequence. Phone durations are calculated by retiming the original phoneme durations corresponding to the visual gestures in the training data such that they sum to the length of the dynamic viseme segment in the video. The audio track is composited with the video for visualization. The lips appear to move in sync with the audio, despite the new word sequence being completely different to the original dialogue. For example, the phrase “clean swatches” can be redubbed with word sequences such as “likes swats”, “then swine”, “need no pots”, “tikes rush” and many others (see Fig. 3). The generated word sequences contain a variable number of phonemes and syllables yet remain visually consistent with the video. This demonstrates the complex relationship between what we hear during speech and what we see. See the supplementary video to see this example and others.

7. DISCUSSION AND FUTURE WORK

This paper describes a method for automatically generating alternative dialogues that synchronize with a video of a person speaking. Dynamic visemes capture the many-to-many mapping of visual to acoustic speech and are a data-driven approach that explain this phenomena. The dynamic visemes corresponding to a speaker’s lip movements are used to construct a graph that describes a sampling of the phoneme strings that could be produced with the articulator motion in the video. A pronounciation dictionary is then used to find the possible word sequences that correspond to each phoneme string, and a language model is used to rank them. An acoustic speech synthesizer generates audio tracks corresponding to the generated word sequences, which can be composited with the original video, producing a synchronous, redubbed video for inspection. The dynamic viseme-to-word search is able to suggest thousands of alternative word sequences for a video, which is far more than if traditional, many-to-one static viseme clusters are used.

An interesting insight of this work is that it highlights the extreme level of ambiguity in visual-only speech recognition.

A sequence of lip motions can legitimately correspond to a vast array of phoneme strings, so recognition is highly dependent on the language model and contextual information. It suggests that better audio-visual language modeling is key to improving recognition accuracy.

Future work will focus on investigating the effect of higher level n-gram language modeling on the grammaticality of the generated sentences, and using different training corpuses to vary the context, style and language of the word sequences.

8. REFERENCES

[1] S. Taylor, M. Mahler, B. Theobald, and I. Matthews, "Dynamic units of visual speech," in *ACM/ Eurographics Symposium on Computer Animation (SCA)*, 2012, pp. 275–284.

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

[3] Janet Jeffers and Margaret Barley, *Speechreading (lipreading)*, Thomas Springfield, IL:, 1971.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] T. J. Hazen, K. Saenko, C. La, and J. 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.

[12] 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.

[13] 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.

[14] J. M. De Martino, L. P. Magalhães, and F. Violaro, "Facial animation based on context-dependent visemes," *Journal of Computers and Graphics*, vol. 30, no. 6, pp. 971 – 980, 2006.

[15] S. Taylor, B. Theobald, and I. Matthews, "The effect of speaking rate on audio and visual speech," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2014.

[16] Carnegie Mellon University, "The CMU pronouncing dictionary [cmudict. 0.7a]," *http://www.speech.cs.cmu.edu/cgi-bin/cmudict*.

[17] ANC, "The Open American National Corpus," *http://www.anc.org/data/oanc/*.

[18] Frederick Jelinek, "Interpolated estimation of markov source parameters from sparse data," *Pattern recognition in practice*, 1980.

[19] Pascal Nocera, Georges Linares, Dominique Massonié, and Loïc Lefort, "Phoneme lattice based a* search algorithm for speech recognition," in *Text, Speech and Dialogue*. Springer, 2002, pp. 301–308.

[20] F. Parke and K. Waters, *Computer Facial Animation*, A K Peters, 1996.

[21] The Centre for Speech Technology Research, "Festival [version 2.1]," *http://www.cstr.ed.ac.uk/projects/festival/*.