A Deep Learning Approach for Generalized Speech Animation Supplemental Material

SARAH TAYLOR, University of East Anglia
TAEHWAN KIM, YISONG YUE, California Institute of Technology
MOSHE MAHLER, JAMES KRAHE, ANASTASIO GARCIA RODRIGUEZ, Disney Research
JESSICA HODGINS, Carnegie Mellon University
IAIN MATTHEWS, Disney Research

ACM Reference format:
Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A Deep Learning Approach for Generalized Speech Animation Supplemental Material. *ACM Trans. Graph.* 36, 4, Article 93 (July 2017), 3 pages.
DOI: 10.1145/3072959.3073699

1 INDICATOR FEATURE SETS

A total of 2379 binary indicator features are used to train the models, where the indicator feature $\delta(\cdot)$ is 1 if the condition is met and 0 otherwise. The indicator features are built around a set of 40 American English phonemes plus *silence*: /sil/, /dh/, /ih/, /s/, /z/, /uh/, /p/, /r/, /aa/, /b/, /l/, /m/, /th/, /ae/, /d/, /g/, /ow/, /k/, /n/, /er/, /iy/, /y/, /w/, /eh/, /ch/, /v/, /t/, /sh/, /f/, /ah/, /sp/, /hh/, /aw/, /oy/, /uw/, /ey/, /ao/, /zh/, /ay/, /jh/, /ng/. This section details the indicator features used in training the models. We use $K_x = 11$.

1.1 Phoneme Duration and Location Indicator Features

Indicator features of this category indicate whether a specific phone span specific consecutive frames. For example, "Does the phone /s/ span frames j through k of the input subsequence?"

- $\delta(x_{i:i+k} == p_j)$ where $i = 1, \ldots, K_x, k = 0, 1, 2, 3$, and phoneme $p_j, j = 1..41 \Rightarrow 41 * (K_x + K_x - 1 + K_x - 2 + K_x - 3) = 1558$ indicator features.

1.2 Articulation and Phoneme Attribute Indicator Features

Indicator features of this category indicate whether a specific phone in a specific location belongs to one of sixty categories describing place and manner of articulation and other phonetic attributes. Table 1 details the attributes and corresponding phoneme sets. This set is largely taken from the 51 phonetic questions from [Odell 1995] (Appendix. B). An example indicator feature in this category is: "Is the phone at the i-th frame a nasal consonant?". An additional indicator feature in this category indicates whether there is a consonant in the first or second half of the input subsequence.

- $\delta(x_i \in PC_j)$ where $i = 1, \ldots, K_x$, PC_j is a phonetic category, and $j = 1, \ldots, 60 \Rightarrow K_x * 60 = 660$ indicator features.
- $\delta(x_{c-5:c-2} \in Consonant \vee x_{c+2:c+5} \in Consonant)$ where x_c is a center frame $\Rightarrow$ 1 indicator feature.

1.3 Phoneme Pair Transitions Indicator Features

These indicator features correspond to an indicator function defining whether a pair of frames correspond to a particular phoneme attribute (vowel or consonant) or data-driven transition categories at a specific location. For the data-driven transition categories, we first collect all AAM parameters corresponding to phone pairs presented in training data and then cluster them to two or three clusters. All details of phoneme pair memberships in each cluster are in the supplementary text file (supplementary.txt). E.g., "Are the phones at k-th and $k+1$-th input frames in a specific cluster of consonant-vowel pairs?"

- $\delta(x_{i,i+1} \in C_j^1)$ where $i = 1, \ldots, K_x - 1$ and $C_j^1, j = 1, 2$ is a cluster of Consonant+Vowel pairs $\Rightarrow K_x - 1 * 2 = 20$ indicator features.
- $\delta(x_{i,i+1} \in C_j^2)$ where $i = 1, \ldots, K_x - 1$ and $C_j^2, j = 1, 2, 3$ is a cluster of Vowel+Consonant pairs $\Rightarrow K_x - 1 * 3 = 30$ indicator features.
- $\delta(x_{i,i+1} \in C_j^3)$ where $i = 1, \ldots, K_x - 1$ and $C_j^3, j = 1, 2, 3$ is a cluster of Consonant+Consonant pairs $\Rightarrow K_x - 1 * 3 = 30$ indicator features.
- $\delta(x_i \in Consonant)\delta(x_{i+1} \in Vowel_p)$ where $i = 1, \ldots, K_x - 1$, and $Vowel_p$ is a vowel starting with phoneme $p \Rightarrow K_x - 1 * 5 = 50$ indicator features.
- $\delta(x_i \in Consonant)\delta(x_{i+1} \in Vowel^p)$ where $i = 1, \ldots, K_x - 1$, and $Vowel^p$ is a vowel ending with phoneme $p \Rightarrow K_x - 1 * 3 = 30$ indicator features.

To examine significance of linguistically-motivated features, we evaluate our approach and decision tree regression on the KB-2k 50 held out test sentences, which share the same feature representation. We compare the results between with full features and only the raw phoneme identity features by measuring squared error in the reference model parameter space, in the raw Active Appearance Model shape space, and in pixel space. Decision tree regression is denoted "Dtree" and Figure 1 shows the results. Note that using only the raw features achieves almost the same performance.

2 PREDICTION AND ERROR PLOTS

To better investigate the performance difference of the various approaches, we evaluate all approaches per each sentence on the KB-2k 50 held out test sentences. We compute squared error in the reference AAM model parameter space, in the predicted shape vertex positions, and in predicted appearance pixel intensities. Decision

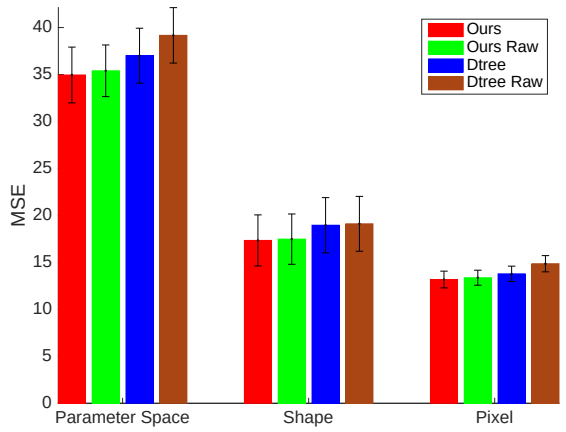

Fig. 1. Showing the mean square error of the KB-2k held out test sentences over the face model parameter space, the full AAM shape space, and pixel space.

tree regression is denoted "Dtree", and dynamic visemes is denoted "DV". In Figure 2, we see that our approach consistently outperforms the baseline approaches on the majority of the held out test sentences.

Figure 3 shows the frame-by-frame predictions of the various approaches for the first face parameter on held out sentence 48. The first face parameter corresponds to how wide open the mouth is. We see that LSTMs and Dynamic Visemes suffer extremely large errors, with the LSTM predictions being particularly jittery. The decision tree regression and HMM-based synthesizer are much more competitive, however occasionally suffers a relatively large error. While the average errors from the decision tree and HMM might be small due to it usually matching the ground truth relatively well, the spikes in error can dramatically reduce the perceived quality of the resulting animation. We observe much smaller error spikes from our approach.

3 RAW USER STUDY RESULTS

KB-2k 50 held out sentences. This is the raw results that were aggregated into Table 1 in the main paper. Each entry corresponds to how many users (out of 25) preferred our approach versus a baseline on a specific test sentence.

Ours vs AAM [13, 11, 12, 15, 11, 19, 9, 13, 12, 16, 10, 19, 14, 18, 14, 9, 13, 11, 7, 17, 12, 12, 14, 18, 19, 11, 7, 15, 12, 13, 12, 15, 12, 14, 13, 11, 15, 9, 16, 9, 9, 15, 10, 17, 13, 14, 14, 11, 10, 14]

Ours vs DV [24, 23, 23, 23, 24, 23, 22, 23, 23, 23, 25, 23, 23, 19, 24, 23, 24, 24, 23, 25, 22, 23, 25, 24, 21, 21, 23, 20, 25, 23, 24, 23, 22, 25, 23, 24, 25, 19, 24, 21, 24, 21, 23, 24, 23, 23, 24, 24, 22, 24]

Ours vs HMM [10, 17, 19, 19, 12, 21, 13, 19, 19, 20, 16, 17, 16, 22, 16, 19, 10, 14, 15, 16, 16, 11, 15, 12, 8, 14, 17, 17, 16, 18, 16, 14, 16, 21, 16, 16, 12, 14, 13, 6, 9, 17, 16, 14, 18, 17, 11, 18, 11, 17,]

Ours vs LSTM [22, 24, 24, 24, 25, 24, 25, 24, 24, 23, 24, 24, 24, 22, 24, 24, 24, 22, 25, 25, 25, 23, 23, 24, 23, 24, 24, 24, 23, 21, 23, 23, 22, 22, 23, 21, 23, 25, 21, 23, 24, 23, 25, 24, 23, 21, 25, 22, 23, 25]

Attribute	Phoneme members
Vowel	/ih/,/uh/,/aa/,/ae/,/ow/,/er/,/iy/,/eh/,/ah/,/aw/,/oy/,/uw/,/ey/,/ao/,/ay/
Vowel starting with /a/	/aa/,/ae/,/ah/,/aw/,/ao/,/ay/
Vowel starting with /e/	/eh/,/ey/
Vowel starting with /i/	/ih/,/iy/
Vowel starting with /o/	/ow/,/oy/
Vowel starting with /u/	/uh/,/uw/
Vowel ending with /h/	/ih/,/uh/,/eh/,/ah/
Vowel ending with /w/	/ow/,/aw/,/uw/
Vowel ending with /y/	/iy/,/oy/,/ey/,/ay/
Plosive	/b/,/d/,/g/,/k/,/p/,/t/
Affricative	/ch/,/jh/
Nasal	/m/,/n/,/ng/
Fricative	/f/,/v/,/th/,/dh/,/s/,/z/,/sh/,/zh/,/hh/
Approximant	/w/,/r/,/y/
Bilabial	/p/,/b/,/m/
Labiodental	/f/,/v/
Dental	/th/,/dh/
Alveolar	/t/,/d/,/n/,/s/,/z/,/l/
PostAlveolar	/ch/,/jh/,/sh/,/zh/,/r/
Velar	/k/,/g/,/ng/,/w/
Unvoiced-Consonant	/p/,/f/,/th/,/t/,/s/,/ch/,/sh/,/k/,/hh/
Voiced-Consonant	/b/,/m/,/v/,/dh/,/d/,/n/,/z/,/l/,/jh/,/zh/,/r/,/y/,/g/,/ng/,/w/
Voiced-Plosive	/b/,/d/,/g/
Unvoiced-Plosive	/p/,/t/,/k/
Voiced-Fricative	/v/,/dh/,/z/,/zh/
Unvoiced-Fricative	/f/,/th/,/s/,/sh/,/hh/
Semi-Consonant	/y/,/w/
Sibilant-Consonant	/ch/,/jh/,/s/,/z/,/sh/,/zh/
Sibilant-Affricate	/ch/,/jh/
Sibilant-Fricative	/s/,/z/,/sh/,/zh/
Front-Vowel	/iy/,/ih/,/en/,/ae/
Central-Vowel	/er/,/ax/,/ah/
Back-Vowel	/uw/,/uh/,/ao/,/aa/,/oh/
Front-Consonant	/b/,/f/,/m/,/p/,/v/,/w/
Central-Consonant	/d/,/dh/,/dx/,/l/,/n/,/r/,/s/,/t/,/th/,/z/,/zh/
Back-Consonant	/ch/,/g/,/hh/,/jh/,/k/,/ng/,/sh/,/y/
Front-Stop	/b/,/p/
Central-Stop	/d/,/t/
Back-Stop	/g/,/k/
Front-Fricative	/f/,/v/
Central-Fricative	/dh/,/s/,/th/,/z/
Back-Fricative	/ch/,/jh/,/sh/,/zh/
Front	/b/,/f/,/m/,/p/,/v/,/w/,/iy/,/ih/,/en/,/ae/
Central	/d/,/dh/,/dx/,/l/,/n/,/r/,/s/,/t/,/th/,/z/,/zh/,/er/,/ax/,/ah/
Back	/ch/,/g/,/hh/,/jh/,/k/,/ng/,/sh/,/y/,/uw/,/uh/,/ao/,/aa/,/oh/
Long-Vowel	/iy/,/er/,/uw/,/ao/,/aa/
Short-Vowel	/ih/,/eh/,/ae/,/ax/,/ah/,/uh/,/oh/
Vowel-Close	/iy/,/ih/,/uw/,/uh/
Vowel-Mid	/eh/,/er/,/ax/,/ao/
Vowel-Open	/ae/,/ah/,/aa/,/oh/
Vowel-Front	/iy/,/ih/,/eh/,/ae/
Vowel-Central	/er/,/ax/,/ah/
Vowel-Back	/uw/,/uh/,/ao/,/aa/,/oh/
Dipthong-Vowel	/ey/,/ay/,/oy/,/ow/,/aw/,/ia/,/ua/,/ea/
Dipthong-Closing	/ey/,/ay/,/oy/,/ow/,/aw/
Dipthong-centring	/ia/,/ua/,/ea/
AVowel	/ay/,/ae/,/aa/,/aw/,/ao/
OVowel	/ao/,/ow/,/oy/,/oh/
UVowel	/ah/,/ax/,/ua/,/uh/,/uw/
silences	/pau/,/h
Total number	**60**

Table 1. Phoneme attributes we exploited for our linguistically motivated indicator feature sets.

Ours vs DTree [13, 12, 12, 13, 15, 17, 13, 12, 13, 19, 17, 13, 16, 16, 16, 14, 17, 12, 11, 17, 13, 18, 13, 14, 16, 17, 11, 19, 15, 10, 12, 10, 14, 15, 16, 17, 14, 10, 15, 12, 13, 13, 11, 20, 16, 17, 17, 14, 17, 16]

Novel Speakers The 24 novel speaker sentences. This is the raw results that were aggregated into Table 2 in the main paper. Each entry corresponds to how many users (out of 25) preferred our approach versus a baseline on a specific test sentence.

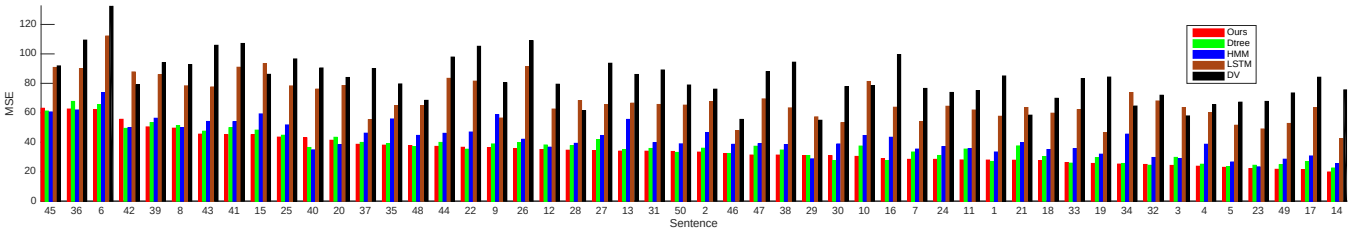

Fig. 2. Showing the squared error (in face model space) for each held out KB-2k held out test sentence. The sentences are sorted in descending order of squared error for our approach. We see that our approach consisently outperforms all baselines.

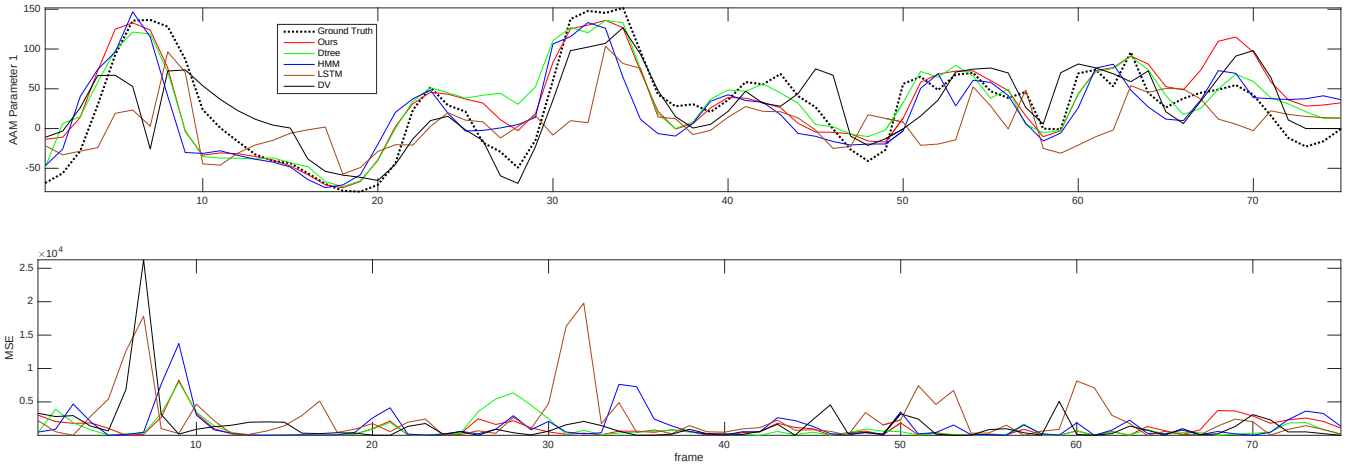

Fig. 3. Comparing the predictions of various predictions for the first face model parameter on held out test sentence 48. The first parameter corresponds to how wide open the moth is. The bottom plot shows the per-frame squared error.

Ours vs DV: [25, 22, 24, 21, 22, 24, 24, 23, 24, 22, 23, 24, 22, 18, 22, 24, 20, 23, 22, 20, 19, 23, 23, 23]

Ours vs HMM [19, 17, 17, 19, 20, 17, 22, 13, 13, 20, 18, 9, 20, 7, 18, 11, 13, 9, 16, 15, 13, 14, 15, 8]

Ours vs LSTM [24, 18, 23, 24, 18, 20, 23, 24, 22, 21, 22, 18, 23, 20, 23, 25, 24, 23, 23, 23, 23, 24, 23, 25]

Ours vs DTree [13, 19, 14, 9, 12, 16, 16, 14, 15, 13, 16, 11, 14, 16, 17, 11, 17, 10, 13, 12, 13, 7, 9, 12]