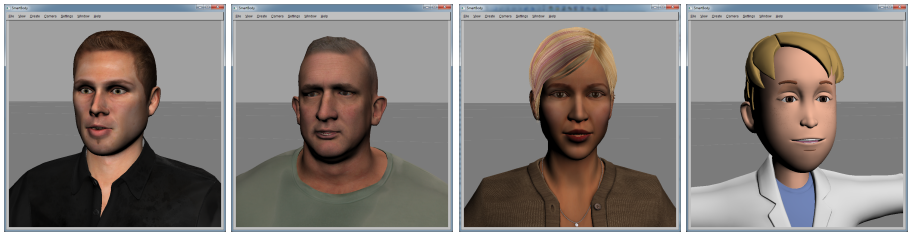
A Simple Method for High Quality Artist-Driven Lip Syncing

Yuyu Xu* Andrew W. Feng† Ari Shapiro‡

Institute for Creative Technologies



Synchronizing the lip and mouth movements naturally along with animation is an important part of convincing 3D character performance. We present a simple, portable and editable lip-synchronization method that works for multiple languages, requires no machine learning, can be constructed by a skilled animator, runs in real time, and can be personalized for each character. Our method associates animation curves designed by an animator on a fixed set of static facial poses, with sequential pairs of phonemes (diphones), and then stitch the diphones together to create a set of curves for the facial poses. Diphone- and triphone-based methods have been explored in various previous works [Deng et al. 2006], often requiring machine learning. However, our experiments have shown that diphones are sufficient for producing high-quality lip syncing, and that longer sequences of phonemes are not necessary. Our experiments have shown that skilled animators can sufficiently generate the data needed for good quality results. Thus our algorithm does not need any specific rules about coarticulation, such as dominance functions [Cohen and Massaro 1993] or language rules. Such rules are implicit within the artist-produced data. In order to produce a tractable set of data, our method reduces the full set of 40 English phonemes to a smaller set of 21, which are then annotated by an animator. Once the full diphone set of animations has been generated, it can be reused for multiple characters. Each additional character requires a small set of eight static poses or blendshapes. In addition, each language requires a new set of diphones, although similar phonemes among languages can share the same diphone curves. We show how to reuse our English diphone set to adapt to a Mandarin diphone set.

Our method fits well within the video game, simulation, film and machinema pipelines. Our data can be regenerated by most professional animators, our stitching and smoothing algorithms are straightforward and simple to implement, and our technique can be ported to new characters by using facial poses consisting of either joint-based faces, blendshape poses, or both. In addition, edits or changes to specific utterances can be readily identified, edited and customized on a per-character basis. By contrast, many machine learning techniques are black box techniques that are difficult to edit in an intuitive way and require large data sets. Based on these characteristics, we believe that our method is well-suited for commercial pipelines. We openly provide our English and Mandarin data sets for download at: http://smartbody.ict.usc.edu/lipsynch/. We present a direct comparison with a popular commercial lip syncing engine, FaceFX, using identical models and face shapes, and present a study that shows that our method received higher scores related to naturalness and accuracy.

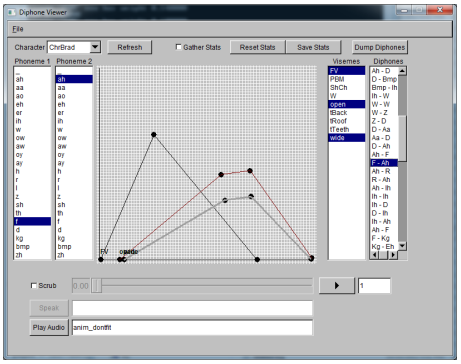


Figure 1: *Curves for F-Ah diphone. The animator selects three facial poses; FV, open and wide, and constructs animation curves over normalized time. The animation can be directly played on the character, or the character could be driven using TTS or recorded audio.*

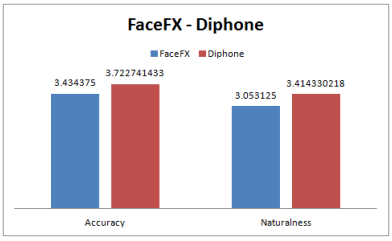


Figure 2: *Comparison between our method (diphone) and FaceFX using both realistic and cartoony characters. We use Amazon Mechanical Turk to collect viewer ratings from about 320 participants.*

References

COHEN, M. M., AND MASSARO, D. W. 1993. Modeling coarticulation in synthetic visual speech. In *Models and Techniques in Computer Animation*, Springer-Verlag, 139–156.

DENG, Z., CHIANG, P.-Y., FOX, P., AND NEUMANN, U. 2006. Animating blendshape faces by cross-mapping motion capture data. In *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, ACM, New York, NY, USA, I3D '06, 43–48.

*e-mail:yxu@ict.usc.edu
†e-mail:feng@ict.usc.edu
‡e-mail:shapiro@ict.usc.edu

I3D 2013, Orlando, FL, March 21 – 23, 2013.