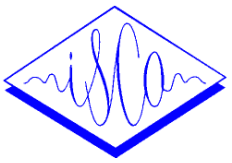


6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A tool for the synchronization of speech and mouth shapes: LIPS

Odile Mella, Dominique Fohr, Laurent Martin[*]*, Andreas Carlen*[*]

{mella,fohr}@loria.fr

LORIA-CNRS & INRIA Lorraine BP 239 F54506 Vandoeuvre France
(*) PROCOMA 74 bis rue des Archives F75003 Paris

ABSTRACT

This paper presents a new approach to improve the phoneme-based lipsync process. The lipsync process is a module in the animation production pipelines of 2D and 3D cartoons. It consists in generating the mouth positions of a cartoon character from dialogue recorded by an actor. The result is a sequence of time markers which indicate the series of mouth shapes to be drawn. We propose to speed up the lipsync process using tools coming from the field of automatic speech recognition. We describe the LIPS tool (LIPS Logiciel Interactif de PostSynchronisation : lipsync Interactive Software) and the generation of the acoustic models required by the tool and we present results obtained on two cartoons. Using the LIPS tool for the post-synchronization of 52 minutes of cartoon reduced the step of post-synchronization from 14 days to 3 days.

1 INTRODUCTION

The lipsync process is a step in the animation production pipelines of 2D and 3D cartoons. It consists in generating the mouth positions of a cartoon character from the dialogue recorded by an actor. The result of this step is a sequence of time markers which indicate the series of mouth shapes to be drawn : for instance an "open" mouth when a "a" is uttered and a "closed" mouth for a "p".
Until now, the lipsync phase (or post-synchronization) has been done by hand: experts listen to the audio tape and write the shapes of the mouth and their timing on an exposure sheet. This traditional method is tedious and time consuming, almost one day for 4 minutes.
In this paper, we propose to speed up the lipsync process using tools coming from the field of automatic speech recognition.
In the section 2, we present the LISP tool. Then, in section 3, we analyze the results obtained on alignment process of two cartoons, one in French, the other in English. Finally, we propose some amelioration to improve the accuracy of the automatic alignment step.

2 LIPS

LIPS (Logiciel Interactif de PostSynchronisation : lipsync Interactive Software) is a tool that, from the speech signal and the orthographic transcription of a dialogue, generates semi-automatically the series of mouths shapes to be drawn. The purpose of LIPS is to reduce the manual operation and then to cut the cost of a cartoon .
With LIPS, the post-synchronization of an utterance is done in four steps:

- The generation of possible pronunciations,
- The signal processing
- The forced alignment with HMM models,
- The manual correction,
- The generation of mouth shapes.

A standard tool of labeling cannot be used because the voices of the cartoon's characters are atypical: shouted voice, child voice, foreign accent, whispered speech, voice of animal characters, laugh...That is the reason why we have modified our labeling tool SALT[1] [1].

2.1 Generation of possible pronunciations

This aim of this step is to provide the main plausible phonetic realizations from the orthographic transcription of the script.
A orthographic transcription can have more than one phonological transcription. First, a word may have several pronunciations. Secondly, the speaker can insert a pause when he takes a breath. Finally, with regard to the French, the "schwa" can be omitted and the speaker can insert a phoneme of liaison. The generator uses a lexicon and some phonological rules to provide a lattice of phonemes.

2.2 Forced alignment

The second part of LIPS performs a forced alignment between all the paths of the lattice and the speech signal.
The alignment, based on the Viterbi algorithm, uses Hidden Markov Models (HMM), more precisely one HMM per phoneme. Each model has 3 states whose the topology is : left-to-right, no skip, self-loop, 6 probability density functions (pdf) with full covariance matrices.
LIPS performs the post-synchronization for French and English cartoons. Regarding the French language, we have trained 35 context-independent models and one more for the pause. As for English, we have trained 48 context-independent models and one more for the pause.

2.3 Preprocessing

The alignment process needs 12MFCC coefficients plus first and second derivatives computed on a signal sampled at 16 kHz. Most soundtracks of cartoons are recorded on DAT with a 32, 44.1 or 48 kHz sampling frequency. So we have to convert the sample rate of the speech signal of the cartoon to adapt it to the models' one. Of course, the signal must first be filtered to avoid aliasing.

2.4 Manual correction

The manual correction of the phonetic alignment is necessary because the forced alignment produced with HMM models may vary in accuracy depending on the quality of the recording and unaccustomed alterations in the pronunciation of dialogues. In

[1] Semi-Automatic Labeling Tool

such cases adjustments may be necessary to match the actual dialogue recording.

To perform the manual correction, we have designed a specific interface where operators can modify the timing of phonemes. End users visualize the audio energy waveforms and can listen to the audio in scrubbing the mouse cursor. The location of phonemes are visualized by labeled markers positioned over the audio in conjunction with the text of the pronounced word. Each phoneme has a start and an end position which can easily be edited by selecting and dragging them with the mouse. Unpronounced phonemes are deleted and additional phonemes or silences may be inserted. Thus, operators have a complete control of the phonetic data and can check that appropriate phonemes have been selected and that their temporal locations are accurate. Energy waveforms have been selected to visualize the audio data due to their familiar representation, and the ability to compute them in real time. Furthermore, senior operators have demonstrated the ability to rapidly position phonemes by observing the shape of the energy curves with minimal audio input.

2.5 Generation of mouth shapes

This a fully automated process which derives 2D and 3D animation data. 2D animation data is generated by matching the phonemes to target mouth position while taking into consideration animation and pronunciation rules as defined by the forefathers of traditional animations. 3D animations data can be derived from 2D animation data and work fine for simple cartoon characters. However, 3D humanoid animation is a much more complex issue which requires the accurate modeling of both human morphology and articulation.

The specific drawings of mouth shapes (the mouth chart) are subject to the discretion of studios and their directors (see an example in figure 1. at the end of this article) . In practice, there are standard mouth charts, however these will vary according to studio specific traditions, individual experiences and desired production aesthetics. In light of the divergent production methods, it is necessary to create a flexible system capable of producing results tailored to the individual needs of each production. SyncMagic recognizes the strong diversity and has addressed two specific issues relative to the issue of lipsync and track reading. Firstly, SyncMagic has automated the tedious task of detecting phonemes and translating these into mouth positions. Secondly, SyncMagic simplifies the integration and usage of track reading data into existing production pipelines. Bellow we will briefly explain some technical issues relative to the process of automatically translating phonetic data to animation mouth code.

The objective of the system is to create the illusion of speech by means of automating the traditional method of manually detecting phonemes and translating these into mouth positions. This tedious process is usually performed by track readers and animators. For each image in the animation, a mouth drawing position will be chosen from the mouth chart. Each image is referenced by a code included in the mouth chart and will inform animator which mouth shape to draw for a specific image. Furthermore, it is often necessary to look within a sequence of pronounced syllables. The goal of this process is to produce an optimal articulation which can be defined as the sequence of mouth movements to be the most believable.

Traditionally, animators will apply a set of animation rules in accordance to the sequence of phonemes being pronounced. For example, the occurrence of a ‘ m b p’ phoneme must necessarily and immediately produce a closed mouth position. Furthermore, simple empiric rule of co-articulation are applied to skip certain phonemes (like short ‘ l’ and ‘r’ sounds) or to add anticipation of mouth positions associated with vowels (such as ‘ ah’ phoneme in str**u**cture. Other phonemes such as the ‘ f’ , ‘ v’ and (often) the ‘l’ sound have dedicated target images which must systematically be visualized, however, it is important that the visual presence of these ‘ special constants’ need to be reduced in time to the strict minimum. The track reader or animator has a set of tools in the form of rules which he or she uses to obtain a sequence of mouth positions based on the succession of syllables. This process allows the animator to focus on a short segments dialogue in which the aesthetics and timing of the articulation can be successively refined.

The choice of mouth positions is critical, particularly when the animator has a reduced set of mouth position (such as five). In such cases it is difficult to produce a smooth animation, discrete changes in the mouth selection have an immediate impact on the aesthetics of the articulation. A simple technique to fluidify the animation consists in applying rules to insert in-between or intermediary mouth positions. This practice involves smoothing the transition between two visually distinct images by means of inserting an image being a mid way (50% or 75%) between the two. Generally, two types of in-betweens are used; firstly, a mouth chart may contain additional mouth positions specifically designed and used for precise visual transitions. Secondly, it is easy to reuse existing visuals in a mouth chart and to apply them as in-between positions. Finally, a more traditional approach used for high quality productions such as in feature film consists in manually drawing in-between position.

The SyncMagic animation process automates the traditional animation techniques for setting in-betweens and obtains results equivalent or superior to the traditional methods. If the number of images allows it, multiple passes can be applied in cascade to obtain an optimal smoothing. Furthermore, the system has a particular affinity to optimize the sequence of images for rapidly pronounced dialogues. SyncMagic articulation technology processes in a similar way and order as is done traditionally by hand. For this, the system defines a temporal window of syllables in which to operate. In this space, a set of hierarchical rules will progressively refine the data to propose an optimal sequence of mouth positions. Three levels of rules exist in this hierarchical process.

The first level, phonetic preprocessing, runs in a phonetic context and will proceed to eliminate phonemes corresponding to simple cases of co articulation and adds anticipation to certain phonemes in a sequence.

The second level operates on phonetic sequences of data produced from the prior stage and produces a series of corresponding mouth positions. A certain number of rules will be applied based on the phonetic pronunciation. The system will first seek to synchronize key positions associated with closed consonants ‘ m p b’ , special consonants ‘ v f l’ , and various levels of priority for different vowels. The set of rules used here can be tweaked to customize the attitude and mood of speech of the animated character.

The third level consists of a set of rules which will operate from both the phonetic and visual context, respectively obtained from levels one and two described above. This stage will put intermediary (in-between) mouth positions where it is visually critical and/or feasible. Visual anomalies produced by level two will be reprocessed and progressively refined until a visually aesthetic result is obtained.

In the event of a rapid pronunciation of dialogue, the system is capable of producing visually satisfactorily results. By doing so the system will prioritize the aesthetics of the animation over the precision. This characteristic correctly simulates the dynamism of movement found in human articulation.

2.6 Mouth shapes vs. phonetic models

We have two choices for designing the system: defining and training one HMM model for each mouth shape or using conventional phonetic models and adding a step for the generation of mouth shape after alignment. We have decided to choose the second solution for two main reasons. First, the system is more flexible: as the number of mouth shapes depend on the cartoon and the character, if acoustic models are used, one does not have to retrain the HHM but only to change the matching rules of the generation module. Secondly, if few HMM mouth models (typically 5 to 9) are trained, each model has to represent very different acoustic shapes thus the models will be less accurate.

3 RESULTS

The system we have described in the previous section is now fully operational on a PC-Linux and used daily to realize the post-synchronization at Procoma company.

3.1 Time sparing

The first evaluation, and the most important from the Procoma point of view, is the amount of time gained: using the LIPS tool for the post-synchronization of a 52 minutes cartoon reduced the step of post-synchronization from 14 days to 3 days.
So, the first version of the software is satisfactory, but the manual correction phase still takes too much time. As it has been described in section 2.4, a major part of the manual correction is due to errors of the automatic alignment. The next sections describe the causes of this misalignment and the possible solutions to improve the labeling part of LIPS.

3.2 Analysis of the alignment

To assess precisely the quality of the alignment process, we have compared the automatically aligned files and the manually corrected ones on two real cartoons production: one in English and one in French.

3.2.1 Results on a French cartoon

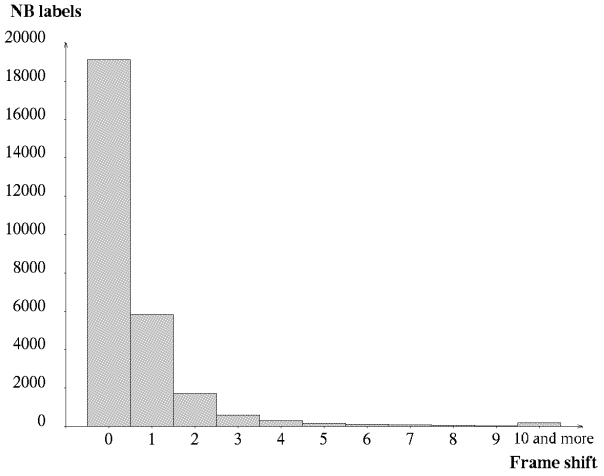
Cartoons contain 25 images per seconds, but for the drawing of the mouth we need only 12.5 images per seconds, and then mouth shapes are duplicated to obtain the 25 images per seconds. So, it is not necessary to get a accuracy better than 80ms (1/12.5 s) *ie* between +40 ms and –40 ms.

Analysis of the boundaries shift

French cartoon	Number	%
Number total of phones	28662	
Insertions	876	3.1%
Deletions	415	1.4%
Confusions	96	0.3%
Correct (-40ms<shift<40ms)	19131	66.7%

The first measure we can give is the number of phone boundaries which shift (absolute value) is lower than 40ms

We can notice than 2/3 of the phone boundaries do not need to be moved. The figure below shows the histogram of the boundaries shift in frames (1frame = 80ms). The mean shift is 0.64 frames.



Analysis of the insertions and deletions

77% of errors are due to the insertion or deletion of the silence model. It is perhaps due to the mismatch between train and test: for the training, the database was recorded in an office environment with light background noise, whereas for the actor's voice, the recording was performed in a high quality studio with no background noise. So, the parameters (pdfs) of the HMM model for silence are not correct and are responsible for the majority of insertions and deletions.

Analysis per character

Character's name	% boundaries correct (shift<40ms)	Average shift (frames)
FRANKLIN	83.41	0.22
MORISSON	82.29	0.44
BIRDSEYE	79.18	0.52
BLAISE	76.34	0.55
APPERT	75.16	0.37
HOMME	73.02	0.43
ETIENNE	72.60	0.37
BIRO	71.27	0.46
DOC-EUREKA	67.58	0.59
JOSEPH	66.94	0.80
BLAISE-PASCAL	65.38	0.55
GRIMOD	65.25	0.65
DRAKE	64.25	0.72
GARNERIN	63.76	0.80
SCHWEITZER	62.63	0.87
EDISON	62.50	0.60
JOUEUR	54.02	0.90
NAISMITH	47.87	0.92
ASSISTANT	46.31	1.03
All together	27.12	4.06

The bad results obtained of the character "Assistant" were surprising. However, after listening to the speech signal, we realized that the actor had taken a very bass voice.

3.2.2 Results on an English cartoon

Analysis of the boundaries shift

English cartoon	Number	%
Number total of phones	13278	
Insertions	282	2.1%
Deletions	1098	8.3%
Confusions	120	0.9%
Correct (-40ms<shift<40ms)	6681	50.3%

The result are worse than those on the French cartoon, half of the boundaries have to be moved during the manual correction phase. The mean shift is 2.5 frames ,that is, 4 times greater than for French. The main difference between the two cartoons is the presence of extreme voices (like animals) in the English one.

Analysis of the insertions and deletions

As for French, the silence model is responsible for the most number of insertions or deletions (32%) and the same explanation (mismatch between train et test background noise) may apply. But there are two other phones that mainly contribute to the insertions-deletions errors: /aa/ and the model for the breath. In this cartoon, there are many laughs (like ah ah ah!) and it is difficult to guess the correct sequence of phones and the quality of the alignment is incorrect.

Analysis per character

Character's name	% boundaries correct (shift<40ms)	Average shift (frames)
OWEN	80.21	0.29
NARRATOR	71.89	0.36
RICKY-TRICKY	67.79	0.81
RICKY	64.6	1.32
OWEN-OWL	63.70	0.56
CONRAD	60.33	1.15
LIZZIE	54.09	1.54
MOLLY	52.87	2.15
MOL-LIZ-FUNG	38.81	1.46
SKIPPER	32.13	3.28
MELVIN	26.59	7.06
VLAD	14.29	9.66

The above table confirm our hypothesis. When the voice are standard, like Owen's or Narrator's voice, the accuracy is quite good, but for extreme voice (like Molly or Skipper) there is a huge degradation.

The results for the two last character of the table (Melvin and Vlad) are awful but these characters do not pronounce a single word, they only cry and laugh. This explain the poor result.

3.3 Conclusion

To summarize the analysis of the results, the main cause of errors are due to extreme voices, laughs, cries, breaths and hesitations. In the next section, we propose some improvements to solve these difficulties.

4 FUTURE IMPROVEMENTS

To take into account atypical voices, we will try two solutions:

- the training of several sets of acoustic models, one for each type of specific voice, for instance bass voice, men, women, , high-pitched voice, voice of animals...
- the adaptation of the acoustic models: using MLLR or MAP algorithm.

To deal with laughs, cries or breaths, we want to create new HMM models for these sounds as soon as we will have a sufficient number of occurrences to train them.

REFERENCE

[1] Odile Mella and Dominique Fohr, "Semi-automatic phonetic labelling of large corpora", Eurospeech'97. (Rhodes, Greece). 1997.

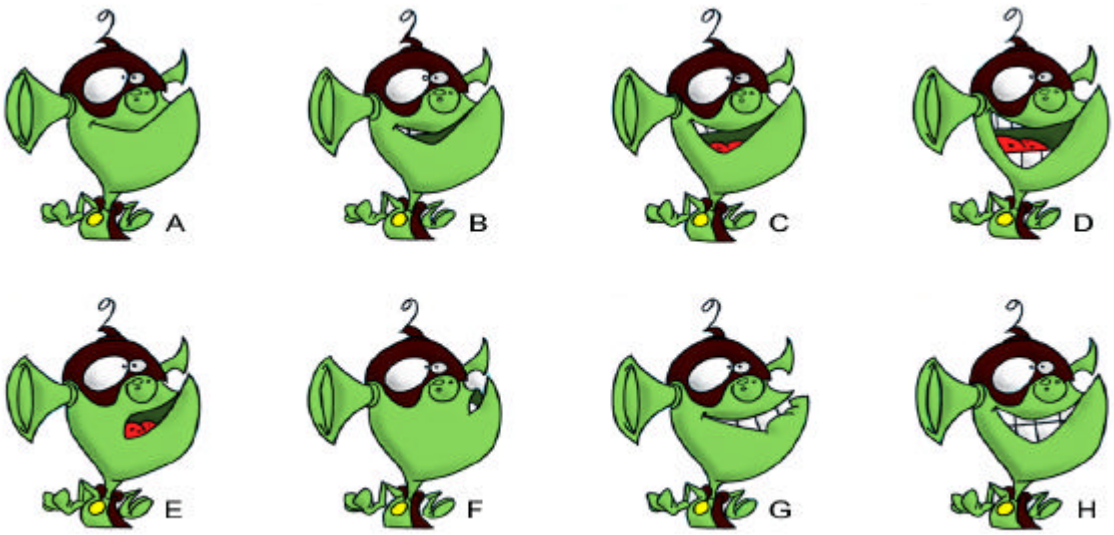


Figure 1. : Example of a set of mouth shapes