Lip Syncronisation Based on Wave Patterns

Pierre Steinmann Bach

Kongens Lyngby 2005
IMM-B.Eng-2005-0

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-B.Eng: ISSN none

Summary

This report contains a short introduction to the basic elements relevant to the topic of lip synchronization as well a general explanation of the idea behind it and attempts to conclude whether it would be possible to implement a fully automated tool for lip synchronization between a arbitrary voice file and an animated head. It then presents 2 different approaches to the speech processing problem, one utilizing the traditional approach of Hidden Markov models and another using Linear Predicative Analysis. A look at the possible implementation of the facial mimics is analyzed as well and in addition to this an evaluation of the products currently on the market is carried out. Based on these things it can be concluded that a fully automated tool for lip synchronization can be implemented and is currently in existence on the market but depending on what the needs are the question of whether an in house implementation should be undertaken or an off the shelf product would be viable does not yield a conclusive answer.

In essence the choice stands between the 2 commercial products LifeStudio:HEAD, Lipsync 2.0 and an in house implementation of the LP based method with a parameterized approach to the facial animation.

Contents

Chapter 1

Introduction

The central goal of this report is to analyze and derive a conclusion as to whether it's possible to fully automate the process of providing a fluent and efficient lip synchronization to an animated head based on an input of an arbitrary sound file with a recording of a voice sequence, in this case in the wav format. In addition to this it would be interesting to take a look at to which extent it would be possible to provide extra features in the sense of facial mimic's based on the information derived from the sound file.

As a general overview such a tool can be described as consisting of the following elements:

- A component that handles the processing and analysis of the signal on the wav file, reduces noise and reduces the sampling to a level where no visual loss in terms of the synchronization would occur but a significant reduction in processing time would be obtained.

- A component that handles the processing of the sound file itself and generates an output that specifies the content of the file with regards to the speech.

- A component that takes the before mentioned output as input and generates the corresponding facial mimics, this being pure lip synchronization, or lip sync combined with other facial mimics.

To obtain a complete overview of the subject and the technology involved a market analysis with regards to what products may exists is carried out as well as a description of the different technologies that can be used upon developing such a tool.

Chapter 2

Project Proposal

For animated film as well as for games there is currently a resource intensive process involved in the task of creating believable facial movement. One of the key elements of believable facial movement is lip movement that is synchronous with the actual speech. In most cases this problem is solved by handling the animation of the mimics manually, or by generating a crude imitation of lip movement that resembles a person speaking but does not correspond with the actual speech. These solutions are either expensive or not very credible.

A possible solution to this could be to create a a fully automated generation of lip movement based on sound.

The goal with this project is as follows:

- To Carry out an analysis of what currently exists on the market today within this particular field.

- To take a closer look at the different options that exists to solve the problem described above. An example could be a solution where phonemes were used as partial interpretation in interpretation of sound.

- To attempt a partial implementation of such a tool based on the analysis and methods covered earlier in the project. Possibly with a graphical representation

in OpenGL

- To take a closer look at additional features, for instance accent, differences based on language, as well as other facial expressions.

The goal is to carry out an analysis of what is currently on the market, take a look at what possibilities exists with the current technology and attempt a partial implementation. Given the duration of the project it is not expected that any extensive implementation will be available.

Chapter 3

Background

Given that the area that the project deals with is of a rather narrow scope a general introduction to the area is provided here. It covers the background information with regards to areas such as how voice is usually generated and how it's classified.

3.1 Physiology

Following background information is based largely on [6] section 2, and [3] chapter 3 part 2.

Since the core of the project deals with generating believable lip movement according to speech and that some of the technologies covered later is tied closely to the general structure of how voice is generated a short introduction to the physiology is in order.

The Vocal tract is what is responsible for generating all sounds produced. The key parts are the glottis and the vocal cord along with the oral and nasal cavities. The flow of air is deliver from the lounges and the diaphragm and when passing through before mentioned parts that regulates it, sound is produced. The glottis centered just above the vocal cord is what regulates the sound production

through contractions which are able to produce the sounds that consonants and vowels consist of. Of the 2 cavities the nasal, which is fixed and thus doesn't change volume is the least important. The oral cavity on the other hand has the ability to change volume. What is more important is that [6], section 2

"the vocal tract can be approximately modeled as a concatenation of a number of cylindrical tubes of uniform cross-sectional area. Though these values have great variation from person to person, the relative distribution is similar for a given vowel"

This is of importance with respect to certain methods of implementing lip synchronization described later.

3.2 Phonemes

Following background information is based largely on [2] and [3], chapter 3 section 3. Phonemes are the smallest identifiable units of sound in the phonetic system. Phonemes are abstract units, in the sense that there's no set amount of phonemes defining a language and it varies greatly from language to language. Each languages has a different phonetic alphabet and combination of phonemes, and a language can consist of between 20 to 60 phonemes. As an example it can be mentioned that there's around 35-40 phonemes in standard American English, depending on the dialect.

In order for people to be able to handle continuous fluent speech the brain carries out pre processing of the following phonemes and thus how a current phoneme is produced physically differs slightly depending on the previous and proceeding phoneme. These variations are called allophones and are a subset of phonemes. While this in itself is not of great importance for the process of lip synchronization they are of some interest since they are the reason for the phenomenon of co articulation, which will be addressed later.

An example of co articulation could be [3], chapter 3 section 3

"a word lice contains a light /l/ and small contains a dark /l/. These l's are the same phoneme but different allophones and have different vocal tract configurations."

In an attempt to create a rough overview of phonemes the following categorizations can be helpful:

- Vowels and Consonants. As with the normal alphabet it makes sense to divide the the phonetic alphabet into 2 groups, but for a different reason. Vowels makes out the cornerstone of voice recognition and lip synchronization since they're easy to detect compared to consonants, mainly due to stability and a considerably higher amplitude. Consonants on the other hand consists of rapid changes which renders them hard to synthesize and thus categorize, something the different methods try to compensate for in various ways.

- Voiced and Unvoiced. These can then be divided into 2 sub groups: Voiced and Unvoiced. If one is to take a strictly phonetic approach to things then both are equally important, but in most cases the unvoiced are of rather little significance with regards to lip synchronization. This is because they won't be visible and could be reduced to the previous phoneme for sake of simplicity. That being said some methods rely on a complete phonetical analysis.

Vowels are always voiced whereas consonants can be both voiced and unvoiced.

- Nasal Cavity and oral cavity. As mentioned in the section on physiology there's 2 cavities that influence the production of sound. This distinction is important because nasal phonemes like the unvoiced, are of little importance to the actual lip synchronization since they're not visible.

- Additional categorizations. Apart from the above mentioned there's a long row of other disquisitions, mainly with regards to the vowels and the way they're produced in the vocal cord. While they're of interest to the actual production of sound and in phonetics in general they're of little relevance for the subject here and thus wont be covered.

As a last thing it's worth mentioning that there's a rather large amount of different phonetic alphabets available, some local, others international and some specially designed for the use of computers. The most well known of these alphabets would be the IPA (International Phonetic Alphabet).

3.3 Syllables

Following background information is based largely on [7], abstract. Syllables needs a short mention because an alternative approach to lip synchronization could be going through syllables. Syllables are the second smallest unit in the phonetical world and by using syllables one would automatically avoid the problems that can occur with co articulation.

3.4 Visemes

Following background information is based directly on [2] Visemes are like the phoneme abstract units, and a set of visemes are visual references to phonemes. Visemes represent the different facial position that is needed to express the lip movement that occurs during speech. While there's between 20-60 phonemes n a language, these correspond to as little as 12, which is the number Disney used, to 18 which is what a deaf person would need for lip movement to be believable.

As mentioned in the section on phonemes, there's a large number of phonemes that can be eliminated and included under other phonemes when it comes to visualizing the speech.

Chapter 4

Analysis

This chapter attempts to take a closer look at some of the different methods which which speech processing can be handled as well as the actual animation process. It's attempted to give a general introduction to that specific approach to the subject and based on that derive what weaknesses and benefits there might be. As it's stated in the introduction the process of lip synchronization essentially consist of 2 modules. The speech processing part, and the mapping part/animation part.

4.1 Analysis of Speech Processing Methods

The difference between voice recognition and the speech processing are fairly small and as a result of this lip synchronization draws heavily on the methods and discoveries made in the field of voice recognition. That also means that traditionally approaches using Hidden Markov Models and the likes, which are heavily utilized in voice recognition, are the most common method of doing lip synchronization. Lip synchronization has a slight but rather important difference. Unlike the voice recognition area it is not essential to identify and classify every phoneme correctly as long as the result would be solid enough to enable a credible lip sync. This means that alternative approaches and methods that

would be considered severely lacking in the field of voice recognition would be a valid approach in this area.

Apart from a general description and evaluation of the different methods there's a few specific things that' be worth noting with regards to what the goal of the project here was.

- It is of rather large relevance whether a method is language independent especially worth regards to the game industry where it'd allow for a much higher degree of localization.

- Most systems requires some degree of training. In most cases initial training is needed, but unless a system is character independent then additional training would be needed whenever there's either variations in the articulation of the speech or when whenever new voices are utilized. In the character independent systems this is considered a learning and classification problem and usually solved by using a neural network.

- The systems that are used depend on a certain amount of preprocessing of the speech signal. This includes establishing a specific frame size for the signal as well as possible noise reduction, pre-emphasis and hamming windowing [1] of the signal

Below are 2 different approaches tot he same problem. The use of hidden markov models is the traditional approach to the issue of lip synchronization while the use of LP analysis is a newer alternative. Other methods exists as well but are not described in detail here. 1 such system could be a syllable based system that relies on language structure to identify the syllables and the corresponding visual syllables, altogether avoid the issues with regards to co articulation and other transitions between the different phonemes that normally causes problems.

4.1.1 Linear Predictive Analysis

This section is based on [6] Linear Predicative Analysis is one of several possible ways of solving the issue with regards to lip synchronization. In the background section it's described how the vocal tract can be approximated by a tube and this is exactly what this method takes advantage of. The reflection coefficients obtained as a by product from the LP analysis are directly correlated to the shape of the vocal tract and can thus be translated into their corresponding phonemes by using a neural network for training. LP analysis has a weakness

[1] Isolation of specific part of the signal to allow for clear analysis of these parts

though, in the sense that it's inadequate for use with consonants and thus will not provide accurate phonetical translation.

More generally the process can be described as consisting of the following parts.

- The LP analysis itself with the reflection coefficients and the neural network to train the coefficients to correspond with the respective phonemes.

- The Energy analysis that covers consonant information and transition between vowels. Consonants has a lower energy level than vowels and can thus be identified and the average energy pr frame of the signal can be recorded. This process is already done as a part of the LP analysis and makes it possible to modulate the already recognized vowel simulating the transition between vowels, a transition that in essence are consonants.

- Zero Crossing. To avoid false closings of the mouth due to such things as unvoiced phonemes the zero crossing rate is used. To detect this the short time average crossing rate which is 49 pr 10 ms for unvoiced and 14 pr 10 ms for voiced speech, is used making it possible to identify in which cases they occur and avoid the problem.

The advantages of this method is that since it's not dependent on any language structure and does not preform speech recognition it's language independent. In addition to that the method is speaker independent and only requires voice input and intimal training of the neural network it relies on.

4.1.2 Hidden Markov Models

This part is in large based on [5], chapter 15 Traditionally seen the use of Hidden Markov Models (HMM) [2] in lip synchronization has long been the standard procedure. This is due to its use in the field of speech recognition since lip synchronization and the use of phonemes to carry out this task is in large parts a subset of the accurate identification of phonemes required in the field of speech recognition.

A HMM based system consists of the following components

- The language model consists of probable word sequences and is often based on the bigram model thus utilizing a first order Markov assumption for word

[2] [5], page 549, "A HMM is a temporal royalistic model in which the state of the process is described by a single discrete random variable"

sequences. The bigram model allows for recognition of a combination of 2 probable words in connection with each other in the sense that if 1 word appears then the probability of the next word appearing is directly correlated to the previous word as well as the next word.

- The word model which consists of all the pronunciation model that compose words.

- The pronunciation model which gives a distribution over phone sequences. In essence the pronunciation model is composed of all the different dialects in which a word could be spoken.

- The phone model which describes how a phone maps into a sequence of frames

The phone model are the elements of the pronunciation model in the sense that the pronunciation model consists of a series of phones that form a word in a "dialect" . The different pronunciation models are then part of the word model in the sense that the different pronunciation models represent a specific word. These are again combined into the language model to compose an entire structure of probeable wordings in a specific language

All the different models are HMM based and as such they can be combined into 1 big HMM. Transitions in this would occur between phone states with a given phone, between phones in a given word, and between the final state of one word and the initial state of another. The transition between words would occur with the probabilities specified in the bigram model.

The problem that occurs with a HMM based model is that the structure of the pronunciation models are hand built. This isn't a problem with regards to widely used languages since there's already large pronunciation dictionaries available but it makes the method very language dependent and requires a lot of work if a new language is to be supported. In addition to that the probabilities of transitions between phones in the phone model needs to be obtained and this is done by using actual speech data that has been hand labeled. This is again a resource heavy process and it imposes the problem that the phone model to some extent is user dependant. The problem has been remedied somewhat since expectation-maximum (EM) algorithms only needs the actual speech data to initialize and then generates a much better set of parameters for the probabilities than even a full set of speech data can provide.

4.1.3 Sub Conclusion with regards to Speech Processing

The HMM methods allows for a more precise phonetical representation of the speech and will under ideal conditions be able to produce an output that delivers a very realistic result. The main problem is that it's language and character dependent and it requires a new pronunciation model built whenever a new language is encountered as well as additional training data for the phones. An LP based method on the other hand is both language and character independent and only requires initial training. Also the LP method augmented with the energy analysis deals efficiently with some of the problems that arise with regards to co articulation and allows for animation of the lip synchronization with expression of how strongly the words are uttered.

With that in mind there's little question with regards to which method would be preferable in the sense that an LP based method in a comparison between these 2 methods is by far superior.

4.2 Analysis of Animation Techniques

This section is largely based on [1] and [2] There are generally 2 approaches to the process of lip synchronization of a face: a parameterized solution or a purely muscle based approach

4.2.1 Parameterized Facial Animation

The standard method of doing facial animation in combination with lip sync is to have a set of parameters that then modify the control points accordingly, creating the wanted expressions.

One of the common ways of doing it is having a phonemes to visemes based system.

The phonemes are grouped into categories that have similar visemes and the input dictates the interpolation between the frames. As mentioned earlier with regards to phonemes and visemes the problem of co articulation occurs and in the straight forward edition it's often needed to manually adjust the viseme interpolation to obtain realistic results as well.

This simple model has a series of flaws. It doesn't account very well for co

articulation which can result in rather and with the basic morphing between the visemes there's no simulation of the strength with which the words are spoken. A large part of these issues can be addressed in the speech processor letting it address some of the problems with co articulation and adapt the phonemes accordingly. Also if the system does not morph directly between the differences visemes but allows for an indication of how strong the respective phoneme was uttered to influence the morphing of control points thus giving it a more natural look.

In general there's a series of smaller issues that has to be addressed when dealing with parametric facial animation but it has been the best solution so far, also given the relative simplicity of it and the low requirement in terms of resources.

4.2.2 Muscle Based Facial Animation

An alternative to parametric based lip synchronization is muscle based facial animation. Muscle based facial animation has been used mainly in the medical industry as well as at very high ends system. The basic idea is to emulate the muscles that exists in the face fully and thus mimic reality. The work required to do such a thing is considerable given that the entire face has to be built based on the muscle groups but it allows for a much finer control of the facial mimics based on the phonemes and thus a much more natural result.

4.2.3 Sub Conclusion with regards to Animation

With regards to the gaming industry at the current time there's little question of what technique has the most relevance. Muscle based facial animation is too resource intensive for the time being and the details obtained from this would to a large extent be lost on a in game character.

The parameterized method coupled with an intensitivity indicator from the speech processing unit would to a large degree suffice and fulfill the needs with respect to current day game industry and makes sense as well since the different standardized multimedia formats now support it eg MPEG 4 [3], chapter 7.

Chapter 5

Market Analysis

5.1 Introduction

As a part of the project a general market analysis has to be preformed, both to establish which products are currently available as well as get a feel of at which stage the development is. There's a wide variety of products available that to fuller or lesser extent fulfill the requirements defined. To limit the extent of the market analysis i have only included the products that i have been able to find that are still commercially available, still supported and are used to a wider extent.

Apart from a general description of the products i have attempted to document the capabilities in a way that allows for a somewhat direct comparison. With regards to the actual quality of the synchronization there's to some extent limitations in the sense that i have been unable to obtain technical specifications with regards as to how the actual interpretation of the speech input is carried out as well as to how the interpretation and actual synchronization is done which limits the ability to draw conclusions with respect to quality.

Upon taking a look at the current market one would realize that there's not a lot of products currently available. I've decided to include those products that seem to be the most used and that are currently still under development.

By that i mean that there's several of companies producing lip sync tools that has since then discontinued the products or simply shut down. In addition to this i's become clear to me that a large part of the current game development companies use in house developed tools [1].

5.2 Products

In an attempt to make it possible to compare the different tools i have taken a closer look at some of their technical specifications as well as a general description of them. It is to be noted that products that would otherwise be interesting has been eliminated if they did not provide any support for 3dStudioMax, Maya or equivalent graphical suites, since these are the main tools used in graphical development for the gaming industry.

5.2.1 Criteria

To get an overview of what the products have to offer certain criteria has been evaluated. Since not all of the products has information available with regards to the different things some things are omitted when it's the case. Also all information with regards to the products has been obtained from their respective websites, in their product descriptions, and in they technical specs if available.

The criteria thats been looked at are the following:

- Type of product Basically states whether the product is a direct plug in for a graphics suite or a stand alone product of a kind.

- Formats supported States the different types of audio and, if relevant, object input it takes.

- Number of phonemes and visemes supported Since there in most cases is a direct correlation between the quality of the lip sync that a given tool is able to preform, and the number of phonemes and visemes that the tool is able to utilize, these informations are included. It's relevant to note that around 20 - 60 phonemes will suffice to cover most languages and that around 12-15 visemes can generally visualize these, as prior stated in the chapter on background information.

[1] Based correspondence with several game companies that shall remain unrevealed here

- Degree of automatization In short describes the degree of automatization, and to what extent the tool covers the entire process.

- Technology used to implement lip synchronization States the methods used for implementing the lip sync if these are available. They're previously covered in the analysis part, where the relevance of each method is covered, as well as highlighting the weaknesses and strengths of the difference methods.

- Price Price if applicable is included for reference. These are quoted directly from the respective company's web site and are guideline prices only.

- Quality The overall evaluation of the quality of the synchronization. This is to a large extent subjective and is largely based on video clips of the synchronization if available unless otherwise stated.

- Others Various other aspects of the product not covered in the earlier categories.

5.2.2 Voice-o-matic

Voice-o-matic is the product of o-matic, who produces a wide range of different graphical tools and plug ins aimed especially at the gaming industry. The product is a plug in for 3d studio MAX and it covers the entire spectrum of file analysis to actual lip synchronization. It's fully automated in the sense that it's wizard based with the only requirement that there's already predefined phoneme to visemes mapping in existence, but does allow for keyframe by keyframe management if that's preferable. With the 40 phonemes supported it'll allow for support with most languages in sufficient detail but it does have a major drawback in the sense that it's unable to produce satisfying results without the use of text input to accompany the speech. Also its not language independent, but does supports several of the major languages. What method that's been utilized with regards to the processing of the sound file is unavailable.

In addition it can be mentioned that a test synchronization has been provided on their website, which according to my own opinion does not deliver highly convincing results.

Priced at around 299 USD

As a final remark it can be noted that their customers include such companies as Blizzard, Electronic Arts and others.

5.2.3 LifeStudio:HEAD

LifeStudio:HEAD is a stand alone graphical suit that deal in animation, skinning, synchronization and other aspects of the creation of heads for characters or avatars. With it comes an import export plugin for integration with maya or 3d Studio MAX. As with voice-o-matic it covers the entire spectrum from speech file to lip synchronization and does so fully automated. A major difference from o-matic is that it comes in a SDK edition including aa fully developed API.

The process of doing a lip sync is wizard based and there's a choice between 2 different audio processing methods. The 1st option is the standard phoneme to visemes based method (likely HMM based) that's language dependent. The 2nd option is language independent and allows for much more extensive control than o-matic as well as manual adjustment of accent by manipulating the wave patterns of the consonants and tracks for control of lip rounding and the likes as well as efficient blending between the facial moves. The last solution could possibly be LP based.

The animations i've seen demonstrated in games using LifeStudio:HEAD has delivered satisfying results.

Priced at around 1700 USD

It can also be noted that companies like Square, Nival and Firelfy Studios amongst others uses LifeStudio.

5.2.4 LipSync 2.0

LipSync 2.0 is a stand alone application that processes a sound file and outputs the result in a standardized format, it being either Softimage/XSI Maya Lightwave MPEG-4 FAPs or MIDI, with the respective timings for the phonemes on that format, thus utilizing the standard for parameterized facial timing already predefined in these formats. The speech analysis is compressed into 12 groupings as pr. the definitions that Disney originally came out with. In addition it can be mentioned that product is both language and speaker independent and given the technical paper [4] that introduces the technology behind it it's safe to conclude that the process used is very close to what's been done in terms of LP.

The quality of the synchronization is also satisfiable, as can be seen on the website.

Priced at around 120 Euro

5.2.5 Magpie Pro

Magpie is a stand alone lip sync and timing program and supports the following formats: wav, aif aiff, mp3, mov, avi for audio and 3ds (3D studio MAX) for objects. It has 2 modes of lip sync. A very simple approximated model based on energy pattern from the speech file and a more advanced edition that relies on phonemes to visemes (likely HMM based). it's major drawback is that it only allows for 8 visemes and it's highly language dependent and currently only supports Portuguese, English, Spanish and French. I allows for speech processing modules to be replaced by own modules as plugins.

Unable to evaluate the quality.

Priced at 250 USD for a single user licence and 3000 USD for a site licence.

5.2.6 Annosoft

Annosoft delivers a Lipsync tool as well as SDK's. The tool is a stand alone application that allows for fully automated lip synchronization based on the HMM model and uses 40 phonemes. It also provides SDK's for integration with the development environment if applicable. The actual mapping from phonemes to visemes is then up to the user, based on annosofts output. As with all HMM methods this is not language independent, but has been modified to be able to handle most cases.

Unable to evaluate the quality.

priced at 500 USD for a single user licence pr year or 3000 USD for a full in house pr game.

It can also be noted that companies like Activision, Electronic Arts, Eutechnyx, Funcom, InXile Entertainment, Lionhead Studios ,Microsoft, Pandemic Studios and others uses Annosoft's SDK.

5.2.7 Sub conclusion with regards to the Market Analysis

Based on the product evaluation made above the following conclusions can be made.

If we are to look at the question of whether it s possible to actually obtain a product thats fully automated then it would be fair to conclude that: yes, it is. Some of the products might need the mapping between phonemes and visemes planned out but all of these things are procedures that needs to done once and for all and then the rest of the process can be automated.

The question of which product to pick would be a choice between LifeStudio, Lipsync and Annosoft. Annosofts product is highly professional and is very geared towards the production environment in terms of in house development but i'd be inclined to stick to the 2 others since Annosoft is based purely on HMM and thus isn't fully character and language independent. Annosoft would be a valid choice in the case that you need more control with the entire process, especially in terms of fine tuning, which Lipsync wont give you.

LifeStudio:HEAD and Lipsync 2.0 are 2 very different products and the choice between them would very much depend on whether you'd be willing to and interested in centering the entire creation of characters around the use of this product, in which case it'd be the best choice, or if one is looking to integrate lip sync into an already existing environment. If the ladder is the case then Lipsync would be the best choice since it has the quality of being language independent and character independent, it delivers a good result and it relies on standardized formats for the output allowing for relatively easy handling of the actual animation of the face.

Chapter 6

Conclusion

To summarize then the goal of this project was to determine whether it was possible to produce a fully automated lip synchronization tool and based on the sub conclusion it can be determined that it is.

The conclusions and recommendations that can be made on basis of this are several and depends partly on what type of tool that would suit one the best.

Based on the market analysis it's clear that if one wish to procure such a product then there's currently 2 products that'd be recommendable. In addition to that it can be recommended that if it's chosen to build your own system from scratch the an LP based approach relying on parameterized Facial Animation, possibly through a standardized format would be the way to go.

If it's assumed that it's not the intention to replace and move the entire production over on a new system then LifeStudio would not be the obvious choice and since in house developed tools do take a nice amount of resources to implement i'd recommend Lipsync 2.0

Taking the goals stated in the project proposal into consideration it can be concluded that an implementation was not attempted nor were there any additional analysis made with regards to how facial mimics and expression of emotions such as anger, grief etc. would be expressed based on what could be read from the

speech file. As a result of that it can be concluded that the project feel somewhat short of it's intended goal.

It can also be concluded that the analysis of speech processing methods should have been more extensive given that there's still a great deal of methods that has not been looked at here that might offer better results.

Future work on this project would clearly entail further research into the different speech processing methods and an attempt to make an implementation based on the LP method unless other more promising methods would surface.

Bibliography

[1] Jeff Lander. Flex your facial animation muscles. *Features, Gamasutra*, 2000.

[2] Jeff Lander. Read my lips: Facial animation techniques. *Features, Gamasutra*, 2000.

[3] Sami Lemmetty. Review of speech synthesis technology. Master's thesis, University of Helsinki, 1999.

[4] R. De Tintis M. Malcangi. Audio based real-time speech animation of embodied conversational agents. In *Gesture-Based Communication in Human-Computer Interaction*, 2003.

[5] Peter Norvig Stuart Russell. *Artificial Intelligence - A modern Approach*. Prentice Hall, 2 edition, 2003.

[6] Nadia Magnenat-Thalmann Sumedha Kshirsagar. Lip synchronization using linear predictive analysis. In *Proceedings of IEEE International Conference on Multimedia and Expo*, 2002.

[7] Nadia Magnenat-Thalmann Sumedha Kshirsagar. Visyllable based speech animation. In *Proceedings of EG2003*, 2003.