Speech driven facial animation using HMMs in Basque

Maider Lehr[1], Andoni Arruti[2], Amalia Ortiz[1], David Oyarzun[1], and Michael Obach[1]

[1] VICOMTech Research Centre
Mikeletegi Pasealekua, 57, 20009, Donostia - San Sebastián, Spain
{mlehr, aortiz, doyarzun, mobach}@vicomtech.es
http://www.vicomtech.es
[2] University of the Basque Country, Signal Processing Group,
Manuel de Lardizabal Pasealekua, 1, 20018, Donostia - San Sebastián, Spain
andoni.arruti@ehu.es

Abstract. Nowadays, the presence of virtual characters is less and less surprising in daily life. However, there is a lack of resources and tools available in the area of visual speech technologies for minority languages. In this paper we present an application to animate in real time virtual characters from live speech in Basque. To get a realistic face animation, the lips must be synchronized with the audio. To accomplish this, we have compared different methods for obtaining the final visemes through HMM based speech recognition techniques. Finally, the implementation of a real prototype has proven the feasibility to obtain a quite natural animation in real time with a minimum amount of training data.

1 Introduction

In human-machine interactive interfaces, in order to obtain a communication as intuitive and comprehensible as possible, there is a clear trend to merge different possibilities of presenting information, in particular, speech and facial animation. In multimedia environments, using both audio and animation is a natural and efficient way of communicating. the information appears more reliable if animation is synchronized with sound. Due to this fact, it is usual to see virtual characters in many aspects of our everyday life.

In order to synchronize the character's lip animation with its speech, the phonetic content is needed, that is to say, phonemes and their duration. If the sound is synthesized from written text, this information is generated by the speech synthesizer. There are many applications that merge these two technologies (animation and speech synthesis) in order to create friendly interfaces [1], [2], [3], [4]. Nevertheless, if the animation is generated from audio, the character has a much more natural appearance because its voice comes from a real person, but we have to perform audio analysis to obtain the phonetic information needed. Most of the research done on this type of application concerns English [5], [6]. The presence of minority languages in the area of virtual characters is

very limited. This is obviously due to the lack of both resources and tailored technology.
In this paper, we study the development of a system capable to produce the suitable data to animate faces from natural voice in Basque using open source technologies. We believe that using open source technologies such as HTK or Sphinx may help diminish the digital gap between majority and minority languages. Specifically, we used HTK Toolkit [7] (based on HMM methods) to discover the best approach to obtain a synchronous animation in real time from speech in Basque, using the minimum amount of resources possible. The output of our speech analysis had to be a match between a set of visemes (visual representation of the phoneme) and phonetic data, corresponding to the lip visualization of the virtual character for each frame of the animation. The final application obtained after different analyses and studies was tested in a prototype. The outcome of this test allowed us to use this application to synchronize the animation with on-line audio in real time.

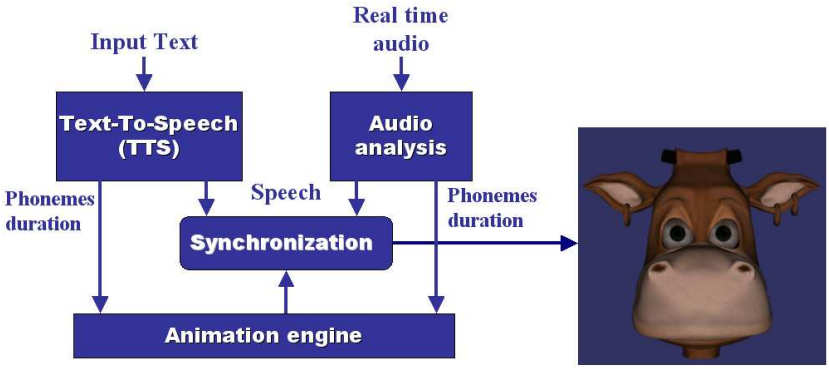

Fig. 1. Animation of the virtual character

2 Technical and Methodological Issues

The purpose of this study is to provide the Basque community with services and facilities not available at the moment. To accomplish this, we have to fulfill the following tasks:

- Research and decide on a method to obtain the suitable data from the speech signal in Basque.
- Synchronize the original voice with the animation from the information provided by the tool developed in the previous analysis.
- Test the performance of these techniques in real-time applications.

2.1 Recognition System

In this part we analyse some approaches to develop the recognition system. First, we describe the methods used and then, we show the results.

Development Approaches. As has been already mentioned, the goal of the project was to obtain the suitable data to animate the lips of the virtual character in real time. To obtain these data, we performed a set of tests using HTK Toolkit. Resources for Basque, such as corpora are scarce and, at present, no open source oral database in the language is available. Therefore, in order to train and test the HMMs, we recorded 250 sentences in Basque, 150 for training and 100 for testing. We added another 50 sentences and used the total amount to create the bigram model. Recordings were done using a Sennheiser desktop microphone (16 kHz/ 16bits/ mono). Recording conditions were not optimal in regards to noise level. The audio files were recorded in a working room with people speaking and with other types of noises, such as steps or street sounds. During training, feature extraction was performed over 25 ms segments every 10 ms. The Basque version of SAMPA was used as phoneme set for the recognizer. Monophone models were created, which consisted of non-emitting start and end states and 3 emitting states (except from the short pause model) using single Gaussian density functions (due to the small amount of training data). The states are connected left-to-right with no skips. For training, we initially set the mean and variance of all the Gaussians of all the models to the global mean and variance of the complete data set. These models were then trained iteratively using the embedded Baum-Welch re-estimation and the Viterbi alignment. The short pause model was added only after a few training cycles. The resulting single Gaussian monophone system was tested using a Viterbi decoder.

Mel Frecuency Cepstral Coefficients (MFCC) vs. Reflection Coefficients (LPREFC). In this environment, we analysed two types of parameterization. One of them was based on MFCCs and the other was based on LPREFCs. MFCCs handle acoustical features of speech sounds and are based on human auditory perception. In this case, the utterances in all data sets were encoded in Mel Frecuency Cepstral Coefficient vectors. Each vector contained the parameterized static vector plus the delta coefficients, as well as the acceleration coefficients. This resulted in 39 dimensional feature vectors. LPREFCs are model based coefficients. We used them to perform the LP analysis. They are closely related to the vocal tract shape. Since the vocal tract shape can be correlated with the phoneme being pronounced, LP analysis can be directly applied to phoneme extraction. 18 reflection coefficients were calculated plus the delta coefficients and the acceleration coefficients. This resulted in 54 dimensional feature vectors.

Complete phoneme set vs. phoneme clusters. For each type of parameterization, we tested two recognition configurations depending on the recognition unit. The first approach used the whole phoneme set to be recognized and a

model was created for each phoneme. The resulting set contained 26 phonemes. In contrast, the second approach consisted on grouping the phonemes on the basis of their visual representation. Since different phonemes share the same visual representation, we obtained a set of 16 models to define, as shown in Table 1[3].

Table 1. Phoneme to viseme mapping

Visemes	Phonemes	Examples
1	p, b, m	apeza, begia, ama
2	d, k, g, rr, r	denda, ekarri, gaia, arrunta, dirua
3	ts, tS, tz	atso, txikia, atzo
4	z, S	zoroa, xoxoa
5	s	hasi
6	f	afaria
7	t	etorri
8	x	ijito
9	n, J	neska, ñabar
10	l	lana
11	L	iluna
12	a	ama
13	e	hemen
14	i	ipar
15	o	oso
16	u	umore

[3] The examples of the table are in unified Basque(Euskara Batua).
apeza: priest/ begia: eye/ ama: mother/
denda: shop/ ekarri: bring/ gaia: topic/ arrunta: common/ dirua: money/
atso: old woman/ txikia: small/ atzo: yesterday/
zoroa: mad/ xoxoa: blackbird/
hasi: start/
afaria: dinner/
etorri: come/
ijito: gipsy/
neska: girl/ ñabar: mixed/
lana: work/
iluna: dark/
hemen: here/
ipar: north/
oso: very/
umore: humour/

Experimental Results. The following table[4] illustrates the results obtained for the two types of parameterization. These results represent the number of visemes recognized by the system. In the case of the phoneme clusters this result is obtained directly. However, in the case of the complete phoneme set we mapped the phone transcriptions of the reference text (the text to recognize) and the text recognized to obtain the viseme recognition rate. The phonemes that have the same visual representation were mapped to the same symbol.

Table 2. Experimental results (viseme recognition rate)

Parametrization	All phoneme set	Phoneme clusters
Mel-Frecuency Cepstrum coefficients	76.45% (Acc:70.76%)	71.32% (Acc:65.28%)
Reflection coefficients	65.33% (Acc: 58.64%)	60.56% (Acc: 54.14%)

Parameterization based on MFCCs. As was mentioned earlier, two different approaches were explored with respect to the recognition module. In the one case, each phoneme was represented by one model. In the other case, each model represented a cluster of phonemes sharing the same visual representation. We compared the results of both approaches and the outcome clearly showed that grouping phonemes on the basis of visemes was not the best approach. The error rate was about 5% worse (the same results were obtained in [9], [10]). This was somehow predictable, since MFCC parameterization is closely related to speech acoustic features. The corresponding MFCC parameterization vectors of the phonemes of the same cluster can be very different from each other.

Parameterization based on LPREFCs. Additionally, another type of parameterization was used to test if another type of parameters was more suitable for grouping the phonemes by their visual representation. In particular, the reflection coefficients were used, since they are not so strongly tied with the acoustic features of the voice. These coefficients use information from the formants and perhaps could be more efficiently correlated with the visual representation of the speech signal. In this case also, we first used the complete set of phonemes to be recognized and, later, generated phoneme clusters. Results show that this parameterization is not suitable either to perform valid recognition based on phoneme clusters. In this case, the results obtained using phoneme clusters were around 4% worse than the results obtained using the complete phoneme set.

MFCC vs. LPREFC. The results obtained using MFCCs were better than the results obtained using LPREFC, both for the configuration of the complete

[4] Acc.(%): represents correct labels, taking into account insertions.

phoneme set and the configuration of phoneme clusters. In both cases the error rate was around 11% worse for LPREFCs. Thus, the use of the parameterization based on LPREFCs did not improve phoneme cluster-based recognition.

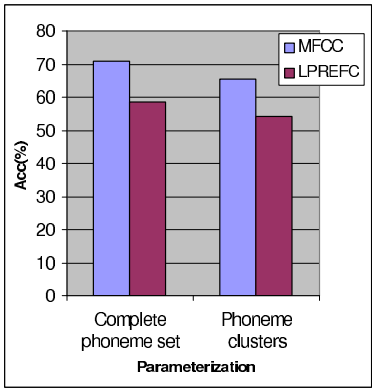

Fig. 2. Experimental results

2.2 Prototype

The application captures the speech signal from the input through a sound card and identifies the appropriate phonemes. As phonemes are recognized, they are mapped to their corresponding visemes. The virtual character is then animated in real time and synchronized with the speaker's voice. The application developed in this paper consists of three modules (Figure 3) (in this paper we concentrate on the component for extracting the mouth shape information from speech signal):

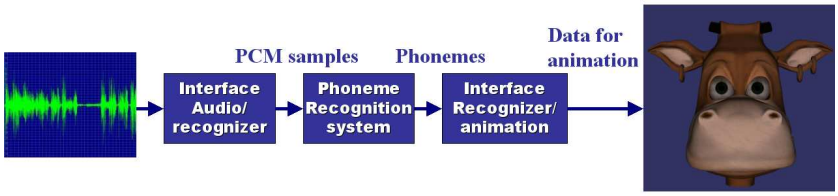

Fig. 3. Diagram of the application developed

– The phoneme recognition system. This module was described in the previous section.

- The module that sends the input audio to the recognition system. Once the off-line recognition system has been developed, it has to be connected with on-line audio, so that the recognizer generates the phonemes corresponding to the speech waveform in real time. To develop this interface we used the ATK API [8].
- The communication interface between the recognition system and the animation platform. Once the phonemes corresponding to the audio samples are recognized, they are sent to the animation module. For this, an interface with sockets was developed, based on the TCP/IP communication protocol. Through this module we fed the animation module with the recognized unit for realistic animation.

It should be noted, though, that due to the lack of sufficient speech data in Basque, the end user must train the system. A minimum of 150 training sentences is needed. In order to facilitate this task, we developed an interface. This interface allows the user to select the sound recording device as well as the format of the audio files.

3 Results and Discussion

Our software fits in today's multi-modal user interactive systems, for which talking characters are an essential part. The present paper focused on obtaining a useful and usable application in the television domain, where virtual presenters are more and more common. The overall aim was to create a quiz type TV program for children using a virtual character. This character is presently running in a popular Basque TV (EiTB) program, in which children answer questions and interact with the said character. In this case, a cow. Results are satisfactory for this first version.

The virtual character is in a virtual environment. The animation reacts to the caller's answers, therefore needs to run in real time. Lip animation runs as the actress who dubs the virtual character in the program speaks into the microphone. This voice is sent in real time to the software developed in this work. The software analyses the stream and generates the data to synchronize the lips with the audio. This information is interpreted by the animation engine.

4 Future work

The results we obtained with the presented recognition application are not accurate enough. However, real time animation obtained is satisfactory.

As a next step, it would be interesting to develop a rich audio-visual database for Basque. With a rich audio database, we could eliminate the training requirement. Besides, with this corpus we could study other possibilities in HTK, such as triphone configuration or a configuration with multiple mixture components. The corpus would also give us information about the structure and grammar of Basque language and more clues as to the reasons for poor performance results.

We could determine if the bad results are due to a lack of data or to another reason. Moreover, if the database contained visual information, it would be possible to perform recognition using both audio and visual features. With a more accurate recognition analysis, we could expand the domain in which the software could be used. It would be integrated in applications such as lip reading for hearing-impaired people or to simply improve comprehension when audio quality is low. Combining audio and animation as means of communication, in noisy environments or when bandwidth is limited, the chances of successful communication increase.

On the other hand, it is interesting to perform a deeper study of the possibilities of the ATK API. We use this API as the interface to communicate live audio with the recognition system. However, its present performance is not as robust as we would like.

We also noted a great dependency between the application and the microphone used. This is another issue that we should address.

5 Acknowledgments

This research is the result of the collaboration in R&D project with Baleuko and Talape (Basque companies in television and film production). We want to express our acknowledgements for the opportunity of evaluating our research in a live diary TV program and also in public events with children.

References

1. Ezzat, T., Poggio, T.: MikeTalk: A Talking Facial Display Based on Morphing Visemes. Proc. Computer Animation Conference, Pennsylvania (1998)
2. Hill, D., Pearce, A., Wyvill, B.: Animating speech: an automated approach using speech synthesis by rules. The Visual Computer **3** (1988) 277–289
3. Magnenat-Thalmann, N., Primeau, E., Thalmann, D.: Abstract muscle action procedures for human face animation. The Visual Computer **3** (1988) 290–297
4. Lewis, J., Parke, F.: Automated lip-synch and speech synthesis for character animation. Proc. CHI87, ACM, New York (Toronto, 1980) 143–147
5. Goldenthal, W., Waters, K., Van Thong, J.M., Glickman, O.: Driving Synthetic Mouth Gestures: Phonetic Recognition for FaceMe. Proc. Eurospeech, Rhodes, Greece (1997)
6. Massaro, D., Beskow, S., Cohen, M., Fry, C., Rodriguez, T.: Picture My Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks. AVSP, Santa Cruz, California (1999)
7. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book. http://htk.eng.cam.ac.uk/
8. Young, S.: The ATK Real-Time API for HTK. http://htk.eng.cam.ac.uk/
9. Lee, S., Yook, D.: Viseme Recognition Experiment Using Context Dependent Hidden Markov Models. IDEAL, Manchester, UK (2002)
10. Dongmei, J., Lei, X., Rongchun, Z., Verhelst, W., Ravyse, I., Sahli, H.: Acoustic viseme modelling for speech driven animation: a case study. MPCA, Leuven, Belgium (2002)