Applying automatic speech generation techniques to audiovisual production

Francesc Alías
Member of the Grup de Recerca en Tecnologies Media at La Salle - Universitat Ramon Llull

falias@salle.url.edu

Joan Claudi Socoró
Member of the Grup de Recerca en Tecnologies Media at La Salle - Universitat Ramon Llull

jclaudi@salle.url.edu

Ignasi Iriondo
Member of the Grup de Recerca en Tecnologies Media at La Salle - Universitat Ramon Llull

iriondo@salle.url.edu

Article received on 01/06/2011 and accepted on 13/12/2011

Abstract
This article presents a summary of the research work of the same title, developed thanks to the grant awarded by the CAC in the VII call of research projects on audiovisual communication. After studying the degree of implementation of speech synthesis systems in Catalonia, we analyze the feasibility of its use for the creation of audiovisual productions. This article presents the findings of the field study and the experiments developed after adapting the speech synthesis system of La Salle (Universitat Ramon Llull) to the Catalan language.

Keywords
Speech synthesis, audiovisual productions, audio description, subjective assessment of quality.

Resum
En aquest article es presenta un resum del treball de recerca que porta el mateix títol, realitzat gràcies a l'ajut concedit pel CAC en la VII convocatòria d'Ajuts a projectes de recerca sobre comunicació audiovisual. Després d'estudiar el grau d'implantació dels sistemes de síntesi de veu a Catalunya, se n'analitza la viabilitat de l'ús en l'àmbit de la creació de produccions audiovisuals. En aquest article es presenten les conclusions de l'estudi de camp realitzat i dels experiments desenvolupats a partir del sistema de síntesi de la parla de La Salle (Universitat Ramon Llull) adaptat al català.

Paraules clau
Síntesi de veu, produccions audiovisuals, audiodescripció, valoració subjectiva de qualitat.

1. Introduction

Speech synthesis is the technique whereby speech is automatically generated with similar characteristics to those of a human voice based on a text input. Speech synthesis systems can be confused with systems that employ recorded voices to reproduce voice messages but it should be clear that, in general, speech synthesis refers to techniques to generate any oral message.

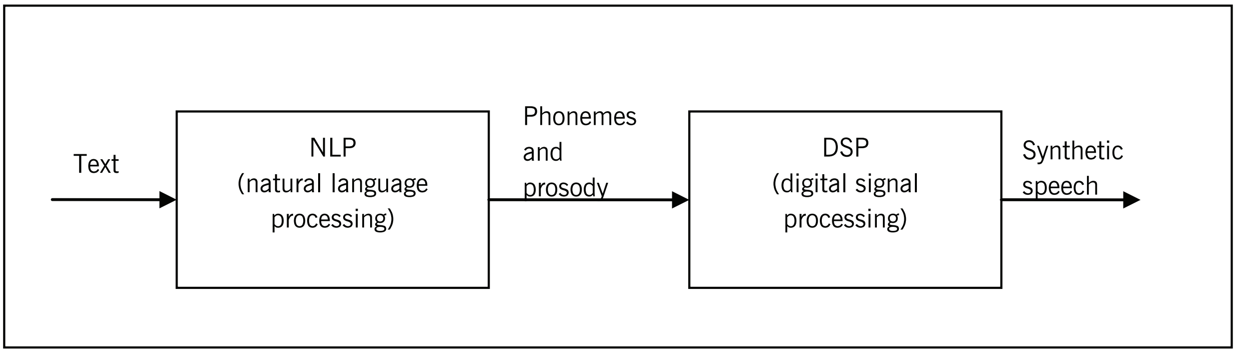
The text input can come from an email, a website or can be written directly on a keyboard. Some of the typical applications of this kind of system are aids for people with a certain disability (for example, visual), a support for language learning, telephone applications, multimedia applications and person-machine interfaces in general.

Far from wishing to imitate the real process used by humans to generate speech, there is a functional model that, employing the resources available today, enables the construction of a system to turn any text input into its corresponding synthetic voice. This functional model, prevalent and widely accepted by the speech technology community, is the one described in the block diagram in figure 1.

As can be seen in figure 1, first we have the natural language processing (NLP) block, whose task is to assign phonetic transcriptions to the words of the text to be "read" (i.e. the sounds that must be produced in the output utterance) and also the associated prosody (how each of these sounds should sound, specifically the characteristics of their intonation and rhythm). Secondly, there is the digital signal processing block (DSP), whose task is to generate the output synthetic speech signal based on the requirements provided by the previous module.

The rest of this article is structured as follows: section 2 presents a review of institutions around the world that stand out for their contribution to the world of speech synthesis. Section 3 presents the most representative findings of a field study on speech synthesis in the audiovisual area in Catalonia, as well as on groups of people with impaired vision. By means of a number of personal interviews, the most relevant opinions have been gathered regarding the degree of maturity reached by this technology, the most significant limitations to its use

Figure 1. Functional model of a text to speech conversion system (TTS)



Source: authors.

and the future challenges in terms of the greater penetration of text to speech conversion in the sectors mentioned. Section 4 describes the process used to adapt the La Salle speech synthesiser (URL) to Catalan, following one of the objectives established in the research project financed. Using this synthesiser, tests have been carried out that are described in section 5, which have subjectively validated the feasibility of using speech synthesis as a tool to generate audiovisual material (specifically examples of advertisements and news items). Finally, section 6 includes the main conclusions of this work and the lines of research that might help to advance towards a higher degree of penetration of speech synthesis in the broadcasting media.

2. Penetration of speech synthesis in the audiovisual world in Catalonia

In order to study the actual degree of penetration of speech synthesis technologies in the audiovisual world in Catalonia, extensive fieldwork has been carried out to gather the opinions of key actors regarding the current penetration and possible future introduction of speech synthesis systems in broadcasting media. Moreover, during this process it was noted that there is a part of the population that are heavy users of speech synthesis systems, namely those with impaired vision. For this reason, this group of users has also been included in the study to discover their opinion regarding the use of speech synthesis technologies within the context of audiovisual productions (Torrens 2010).

Below is a representative review of the most relevant companies, research centres and products related to synthetic voice generation in Catalan. Within this context, companies are included both within the Catalan and international sphere, as well as products available online.

2.1 Universities and research centres

1) TALP (Tecnologies i Aplicacions del Llenguatge i la Parla) from the Universitat Politècnica de Catalunya

Regarding Catalan speech synthesis we must point out, on the one hand, that the TALP research group has its own text to speech conversion system called OGMIOS (*http://www.talp.cat/ttsdemo/index.php*), and, on the other, that it has worked on incorporating Catalan within the Festival platform, developed for the Linux operating system (*http://www.cstr.ed.ac.uk/projects/festival/*), and the result is FestCat, which was included in the Catalan government's Linkat distribution. All these applications can be downloaded free of charge from the FestCat website and are published under LGPL licence terms. For more information, see the website <http://gps-tsc.upc.es/veu/festcat/>.

Some of this work was carried out as part of the Tecnoparla project (speech technologies in Catalan), focusing on the feasibility of voice translation applied to the translation of audiovisual news. The project studied the different key technologies involved in a voice translation system (recognition, translation and speech synthesis), focusing on the incorporation of Catalan, and dealt with the progress made in the three technologies in question and their integration. With regard to speech synthesis in particular, the Festival platform's open programming system was used (Linux) adapted to Catalan (FestCat). You can find more information on the following website:
<http://www.talp.cat/tecnoparla/>

2) GTM (Grup de Recerca en Tecnologies Mèdia), La Salle - Universitat Ramon Llull

This group has extensive experience in the world of synthetic speech generation. Since it was first set up at the end of the 1980s, it has focused on research into Catalan speech synthesis, via work by Martí (1985) and Camps (1992), later continued by Guaus and Iriondo (2000) and Iriondo *et al* (2004), the latter work being focused on expressive (emotive) synthesis in Catalan.

For more information, visit the website:
<http://www.salle.url.edu/portal/departaments/home-depts-DTM-projectes-PM?cf_seccio=PM&pag=1>

3) Barcelona Media - Centre d'Innovació de la Fundació Barcelona Media

Barcelona Media includes a line of voice and language research that investigates language processing, both written and oral, and develops applications in automatic correction and translation, data analysis and processing, automatic text generation from databases and speech synthesis, to achieve tools for the automatic processing of linguistic content in multilingual environments or where human language is the priority means of interaction.

In the area of speech synthesis, their aim is to create a synthetic voice in Catalan, one in Spanish and one bilingual (Catalan and Spanish), as well as making this more natural, expressive and with better intonation (prosodic) and to facilitate the creation of specialist speakers. You can find more information on the following website:

<http://www.barcelonamedia.org/linies/7/en>.

2.2 Companies

1) Verbio

Barcelona-based company dedicated to selling speech technology products.

- In terms of speech synthesis, they offer text to speech conversion in different languages.
  <http://www.verbio.com/webverbio3/html/productes.php?id=1>
- Demonstrations of voices in Catalan: Meritxell and Oriol.
  <http://www.verbio.com/webverbio3/html/demos_ttsonline.php>
- Demonstrations of news:
  <http://www.verbio.com/webverbio3/html/demos_news.php>
  Link on Vilaweb.cat, but it indicates that there are no news items available.

2) Loquendo

Company dedicated to selling speech technology products. With regard to speech synthesis, they offer text to speech conversion in different languages.

- This is a speech synthesis system based on unit selection.
  <http://www.loquendo.com/es/technology/tts.htm>
- Demonstrations of voices in Catalan: Montserrat and Jordi.
  <http://www.loquendo.com/es/demos/demo_tts.htm>

3) CereProc

The company CereProc, in collaboration with Barcelona Media, has developed a bilingual female speech synthesis system in Catalan and Spanish. They offer a synthetic female voice, bilingual in Catalan and Spanish, with natural intonation, available for many different applications. This project has been supported by the Catalan government.

<http://www.cereproc.com/products/voices>

4) Nuance

Nuance Vocalizer (formerly RealSpeak) has a female voice in Catalan (Núria). However, there isn't much information on the company's website.

<http://www.nuance.es/realspeak/>

<http://www.nuance.com/for-business/by-solution/contact-center-customer-care/cccc-solutions>

<services/vocalizer/vocalizer-languages/index.htm>

5) Telefónica I+D

Has a multilingual text to speech conversion system (Rodríguez *et al.* 2008). We have not been able to confirm whether this is an independent product offered by the company (see *http://www.tid.es*). However, the company has incorporated this technology in some of its products, such as the short message reader (*http://saladeprensa.telefonica.es/documentos/24moviles.pdf*) or to help disabled people (*http://saladeprensa.telefonica.es/documentos/22comunicador.pdf*).

2.3 Other products

1) eSpeak

eSpeak is a synthesis system based on formats that work on the Linux and Windows platforms and that can be used under a GNU *General Public License* (freeware).

<http://espeak.sourceforge.net/>

2) JAWS (*Job Access With Speech*)

Aimed at the blind and visually impaired.

- Reads the content of a screen using a synthetic voice.
  <http://www.freedomscientific.com/products/fs/jaws-product-page.asp>
- Includes a voice in Catalan because it includes synthesis systems from other companies, such as Nuance (Núria).
  <http://www.freedomscientific.com/downloads/jaws/JAWS10-whats-new.asp>

3. Penetration of speech synthesis in the audiovisual world in Catalonia

In order to study the actual degree of penetration of speech synthesis technologies in the audiovisual world in Catalonia, extensive fieldwork has been carried out to gather the opinions of key actors regarding the current penetration and possible future introduction of speech synthesis systems in broadcasting media. Moreover, during this process it was noted that there is a part of the population that are heavy users of speech synthesis systems, namely those with impaired vision. For this reason, this group of users has also been included in the study carried out to discover their opinion regarding the use of speech synthesis technologies within the context of audiovisual productions. The details of the fieldwork can be found in Torrens (2010).

Below is an analysis of the findings from the fieldwork based on the different answers obtained from the interviews held with key actors in the sector (radio and television broadcasters, production houses and sound and dubbing studios), by means of interviews held with people working in this sector, both from a technical and non-technical view.

Moreover, a group of users was also interviewed who are potentially very interested in speech synthesis being incorporated in the world of audiovisual communication, such as the visually impaired. The findings of the study will then be presented, contextualised by this sector of society.

3.1 The media

The interviews held with the broadcasting media have been broken down into three large groups: 1) radio, 2) television and TV producers, and 3) sound, dubbing and post-production studios. Most of the leading organisations in the sector were contacted within Catalonia, as the study focused on the application of speech synthesis in Catalan. Out of all these organisations, the following 19 entities attended the interviews via a representative from their technical/broadcasting departments (see Torrens 2010 for more details):

1. Radio: Catalunya Ràdio, 40 Principales Barcelona, COMRàdio, RAC1 and Onda Rambla - Punto Radio.
2. Television and production houses: TV3, 8tv, RAC105tv and Gestmusic.
3. Sound, dubbing and post-production studios: Oido (*www.oido.net*), INFINIA (*www.infinia.es*), Onda Estudios (*www.ondaestudios.com*), Cyo Studios (*www.cyostudios.com*), Dubbing Films (*www.dubbingfilms.com*), Tatudec (*www.tatudec.com*), Dvmusic (*www.dv-music.es*), Seimar RLM Estudios, Soundub (*www.abaira.es*) and Sonygraf (*www.sonygraf.com*).

The following conclusions can be reached from these interviews:

- Both the radio and television broadcasters and the sound studios are aware of the technology of speech synthesis systems.
- Analysing the first of the groups, none of the radio broadcasters contacted uses speech synthesis systems apart from a couple that have only used them to generate a robotic voice or to create a specific effect, and they have not done this using freeware.

There are various opinions regarding the use of speech synthesis technologies in the future: two of the people representing the technical departments of broadcasters believe that they might be useful but only in a complementary way; i.e. to create effects or for automated broadcasters. Another states that the charm and magic provided by a medium such as radio would be lost; the two remaining people think that synthesisers are still far from being used due to the lack of natural expression and intonation in the voice.

- None of the television broadcasters or the production houses contacted uses speech synthesis systems to generate audiovisual products. However, the opinion of the technicians consulted is quite varied. In one case, they state that they are not interested because what people like is a human voice. Others state that they could be used in automatic programmes providing information on the stock market or the weather and also in advertising, documentaries and promotions, due to the great financial savings the generation of these products would suppose. This last statement has been taken from the interview held with the technical representative of the audio department of the TV production house, Gestmusic. Although some technicians believe it is feasible to use a synthetic voice in various applications, they also think that speech synthesis systems have yet to mature to a sufficient level of naturalness to be able to produce a range of intonations (voices that are high or low pitched, young, serious, etc.).
- Only two of the technical departments from the last group (sound, dubbing and post-production studios) have ever used a voice synthesiser, and this only to create music effects or to manipulate voices. The general opinion regarding the penetration of these communication systems in the future is very similar among all the studios consulted. The vast majority of the people interviewed stress that, until speech synthesis systems are perfected further (in the sense of increasing the naturalness of the synthetic voice generated to be able to convey emotions realistically, as a person does), the synthetic voice can't be used, not in the television or radio industry.

As an overall assessment of the idea of introducing speech synthesis systems in the broadcasting media, it can be said that the opinions of the 19 technicians interviewed, in principle opposing the integration of these systems in the process of creating audiovisual content, might change if the emotions of the voice could be synthesised naturally and less robotic synthetic voices could be achieved, thereby being closer to the natural voice produced by humans.

3.2. Potential users

Regarding the interviews held in the context of technologies for the visually impaired, the interviews were carried out with two different profiles: 1) technicians working in the broadcasting media, contained in the previous section, to know their opinion on the use of synthetic voices for audio description (a technology they are already familiar with), and 2) the segment of the population that suffers some kind of visual impairment, as it is essential to take their opinion into account regarding the viability of introducing artificial voices in these media.

Most of the people working with sound technologies (including technicians from the radio, television and sound, dubbing and post-production studios interviewed) believe that synthetic voices could be applied to audio description if they were more natural and "believable" although, in some cases, they also think that it wouldn't save much time and that natural voices aren't worth replacing. Specifically, some believe that speech synthesis systems would have to improve a lot in terms of quality and it has even been stated that it's quicker to record with a person. However, out of all the interviews, there are two that stand out in particular as clearly different from the rest. Specifically, the following has been stated:

- Before incorporating speech synthesis technologies in audiovisual production, we should ask the visually impaired, who are really the end users, regarding the viability of using synthetic voices for audio description and, if they don't like it, this option should be discarded.
- It's always better if the message's intonation and naturalness are good but cost could be a key factor. In this respect, although a synthetic voice may not be completely natural, it could reduce the cost of creating the audio and therefore be more profitable than hiring a narrator.

3.3. Visually impaired users

By collaborating with ONCE [Spanish association for the blind], we were able to interview 51 visually impaired people from all over Spain. The distribution by region of the people consulted is as follows: 16 people from Madrid, 8 from Andalusia, 5 from the Community of Valencia, 3 from Catalonia, 3 from Galicia, 3 from the Balearic Islands, 3 from Asturias, 2 from the Canary Islands and one representative from each of the remaining regions (Cantabria, Basque Country, Castile & Leon, etc.). The breakdown by profession is as follows: 10 retired, 9 lottery ticket sellers, 7 physiotherapists, 2 students, 2 teachers and a single representative for the remaining professions (programmer, musician, speech therapist, journalist, lawyer, office worker, telephonist, bank clerk, etc.). The interviews have therefore attempted to cover a wide range of profiles of potential users with a visual impairment (see Torrens 2010 for a detailed list).

Two quite interesting ideas can be extracted related to broadcasting media out of the interviews held with either totally or partially visually impaired people:

- Almost everyone who collaborated by completing the questionnaire believe that synthetic voices could be used in the future for audio description on television and in the cinema. They say that it would be very useful for a voice to explain everything they can't see in television programmes, documentaries, films, etc. Anything that helps them to lead a normal life and join in the consumption of audiovisual products is welcome.
- Opinions differ regarding the introduction of speech synthesis systems on the radio. More than half believe that it's unnecessary and prefer a human voice. Out of the rest of the interviewees, some believe that it could be useful, depending on the quality of the synthetic voice and others, although they accept it, don't believe it to be essential.

Finally, we can conclude that, once naturalness and emotion is achieved in synthetic voices, audio description could be a good way to progressively introduce speech synthesis systems in the world of audiovisual productions, as almost all visually impaired use these systems. While this advance is expected in voices, the viability of introducing speech synthesis systems on the radio or television seems complex but there is the option of using them in sectors or applications in which expressiveness is not required or when a robotic voice is the aim.

4. Adapting the La Salle synthesis system to Catalan

In the technical area of the project, one of the key phases has been the development of linguistic and signal processing resources to create voices in Catalan. The linguistic resources, such as the phonetic transcription system, the morphological-syntactic analyser, etc. that form part of the natural language processing (NLP) module of the synthesis system have been developed by the La Salle research group over the last few years of study. However, the speech synthesis databases in Catalan are public and have been developed by the TALP research group of the Universitat Politècnica de Catalunya, with funding from the Catalan government, as part of the FestCat project (*http://gps-tsc.upc.es/veu/festcat*).

From this project, two voices have been chosen – Ona and Pau – that are more widespread, given that the speech synthesis system of the Grup de Recerca en Tecnologies Mèdia (Media Technologies Research Group) of La Salle (URL) is based on unit selection and following the pre-established parameters of the prosodic model.

Once there are voice files, the synthesis system has to "create a new voice", i.e. process voice samples so that they can be used to generate a synthetic voice. The creation of a new voice consists of three main stages:

1. Segmenting the database into units of synthesis that determine the start and end of each acoustic unit (diphones, in this case) that make up the messages recorded in the voice files.
2. Indexing and parameterising the database to generate the files in XML format that contain the parameters describing the acoustic content of the database (duration, energy, underlying frequency of the units). At the same time, the cost function selection has to be adjusted, something that involves, on the one hand, pre-calculating all the costs of the database units and, on the other, adjusting the weights of the cost function (Alías *et al.* 2011).
3. Training the prosodic model to determine the most appropriate pronunciation of a text input, to be synthesised by extracting prosodic patterns taken from available voice samples (Iriondo *et al.* 2007).

Once these three stages have been completed, the voices Ona and Pau can be used, integrated within the La Salle speech synthesis system, to carry out the experiments that aim to analyse the viability of using speech synthesis in audiovisual productions, described below.

5. Experiments and results

Different characteristics can be evaluated in the area of speech synthesis, such as intelligibility, naturalness and expressiveness. In some applications, for example in speaking machines for the blind, the intelligibility of speech at high speed is more important than its naturalness (Llisterri *et al.* 1993). On the other hand, correct prosody and great naturalness are essential in most multimedia applications. The evaluation can be carried out at different levels (segment, word, sentence or paragraph) and with different kinds of tests (Campbell 2007).

In order to achieve a subjective evaluation of the viability of using speech synthesis to generate audiovisual material, two perceptive tests were prepared: one using an advertisement and the other a news item. For each test, a set of pairs of stimuli was prepared. Each pair had the same verbal content but one was generated by the synthesis system and the other was read by a person. Once the stimuli were prepared, the most suitable type of test was decided to present them to the listeners, as well as the methodology used to evaluate these stimuli. In the case of the advertisements, only the audio channel was used, while in the case of the news items there were videos with images related to the news and an audio channel made up of a background soundtrack (music, street noise, voices, etc.) superimposed over the track of the narrator's voice.

As has already been mentioned, the aim of the experiment was to evaluate speech synthesis in advertisements and news items. There was a pair of audio files (advertisements) and video files (news items) for each element to be evaluated. Different kinds of presentations were proposed for the stimuli (individually or in pairs) and scales of ranking. Based on recommendation P.800 of the International Telecommunication Union (ITU) (ITU-T 1996), the Comparison Mean Opinion Score (CMOS) was chosen, which allows the comparison of two stimuli, *A* and *B*, as:

- A much better than B
- A better than B
- A slightly better than B
- About the same
- B slightly better than A
- B better than A
- B much better than A

Listeners used this scale to compare and rate the two stimuli presented, listening to them as often as required.

5.1. Advertisements

To evaluate the use of speech synthesis in real situations, a test was prepared with seven advertisements. Two sound files were generated for each advertisement, one based on an amateur narrator reading an advertisement and the other using our speech synthesiser in Catalan.

The test was carried out using the online platform TRUE (Testing platfoRm for mUltimedia Evaluation) (Planet *et al*. 2008), which allows the test to be designed and carried out remotely.

For each pair of audios associated with the same advert, the test participant was asked two questions:

1. "The following audios (*A* top, *B* bottom) correspond to two readings of advertisements. The question is not to evaluate whether you prefer the voice of one woman to another but to indicate your preference regarding their use in advertising, focusing on the NATURALNESS of the pronunciation and intonation:"
2. "With regard to INTELLIGIBILITY, what do you think?"

The test was carried out with 25 listeners (12 women and 13 men) aged between 18 and 66.

The preferences obtained with this test are shown in figure 2, where *A* represents the natural voice and *B* the synthesised voice. As expected, the results show a clear preference for the natural voice, especially in terms of naturalness, although this difference is not so great in intelligibility.

5.2. News videos

The aim of this experiment was to add two customary components to the voice in audiovisual material: an image and a soundtrack in addition to the voice track. A test was prepared with three pairs of news items. Using material taken from YouTube and a voice generated by our synthesiser, news videos were made with three tracks: the video track itself and two audio tracks (background sounds and a voice).

The test was also carried out using the TRUE platform and there was a CMOS with seven categories. 20 people took part (17 men and 3 women) aged between 24 and 41. The users were not informed of the origin of both voices. At the end of the test, participants were asked their sex and age and whether they were speech technology experts. They were then asked two open-ended questions:

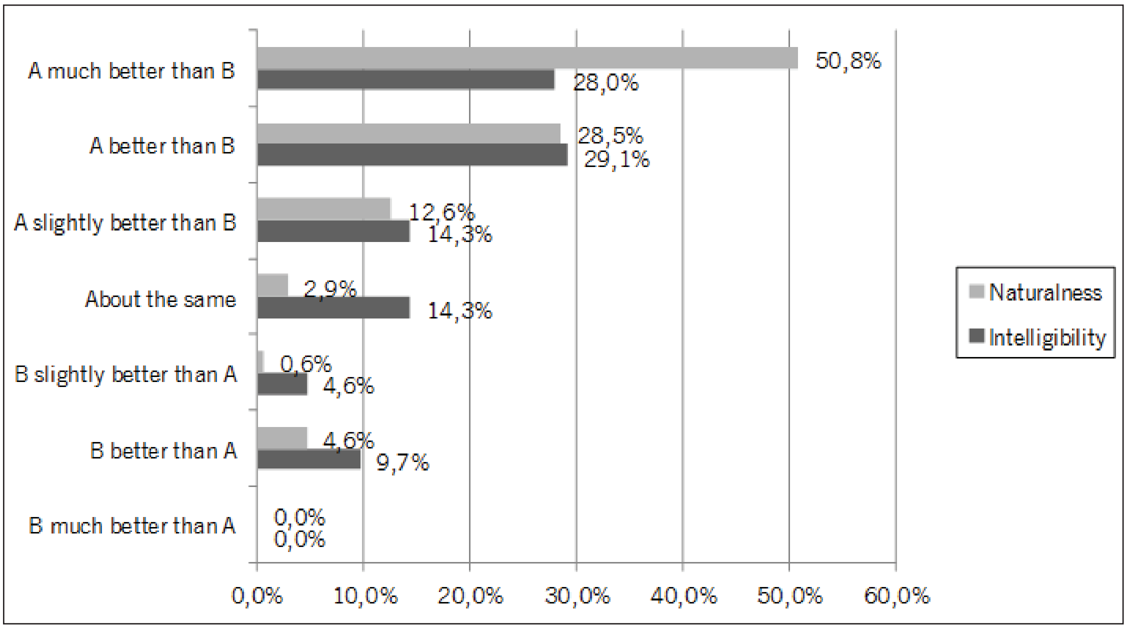
1. "The bottom video voice was generated by a computer. What did you think of it?"
2. "Do you think it's feasible to use speech synthesis to explain news on programmes generated automatically?"

The results obtained are shown in figure 3, where we can see that the majority response is that the natural voice is slightly better than the synthetic one (46.3%). It is important to note that practically 26% of the answers (18.5 % of *about the same* plus 7.4% of *the synthetic voice is slightly better than the natural voice*) indicate that the synthetic voice is acceptable in this context.

If we analyse the responses of the participants where they have given their opinion, after having done the test, regarding the use of speech synthesis to generate news, we can note two general ideas. Firstly, listeners are highly sensitive to certain errors in a specific part of the text, and expressiveness and rhythm need to be improved. Secondly, the majority opinion is that it is viable to use this technology to generate breaking news, for example for a website or semi-automatically generated programmes. To illustrate these two conclusions, below is a broad collection of the responses obtained in tables 1 and 2.

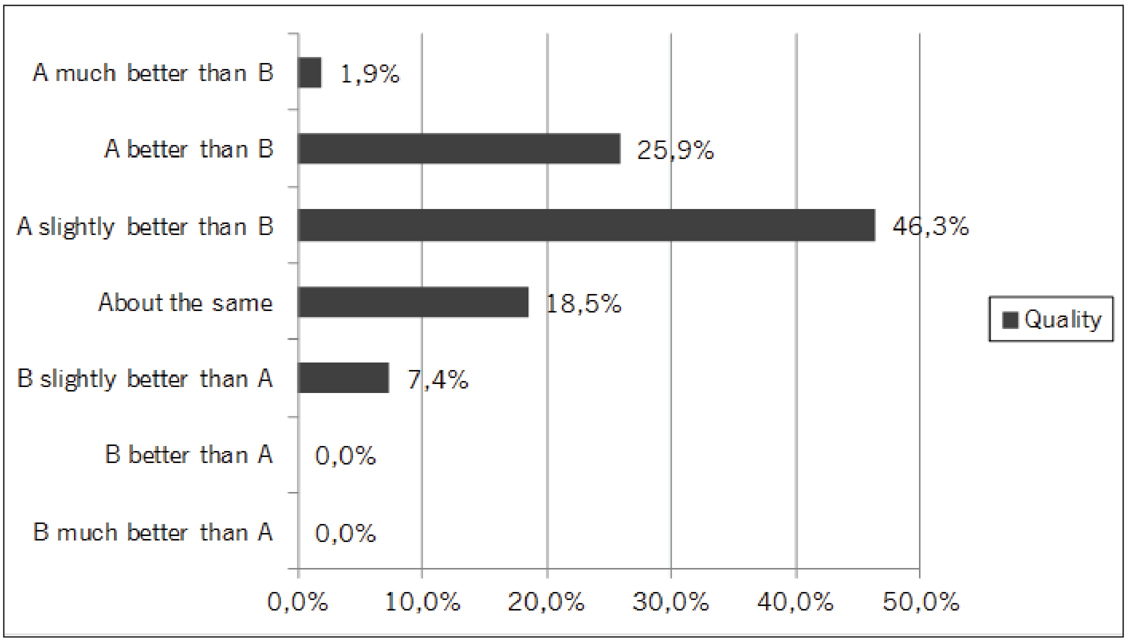
If we compare the results with the advertisement test we can see that adding video and background sound helps to dissimulate the errors of the synthesis and to divert attention, making the use of a synthetic voice more acceptable.

Figure 2. Results of the advertisement test regarding intelligibility and naturalness. A corresponds to the natural voice and B to the synthesised voice



Source: authors.

Figure 3. Results of the news videos test regarding the quality of the voice in off. A corresponds to the natural voice and B to the synthesised voice



Source: authors.

Table 1. Selection of answers to the question "The bottom video voice was generated by a computer. What did you think of it?"

"Quite acceptable, although a bit slow and with some mistakes in specific sounds."
"Good quality in general, although there are some gaps and jumps in the intonation."
"Quite good but you realise it's not human at certain times."
"Sometimes very good (even better than the original), and others not. The "false notes" from time to time bring down the overall quality."
"Not very natural, although you could notice a little expressiveness and the audio's quality was quite good. Maybe problems in keeping a constant rhythm; you notice jumps in the rhythm."
"The voice is a little metallic. The intonation isn't natural enough. Without doubt, you realise a machine is talking to you all the time. Nevertheless, the message is understood correctly."
"Quite good, especially in the first one. The background noise dissimulates the errors. Depending on the subject, the narrator's style should vary (e.g. in a festive atmosphere, speaking more quickly)."
"You notice that it's a synthetic voice but it doesn't bother you because it goes with with the music and images and its quality means you can understand clearly everything it says, even better, sometimes, than the real one."
"Quite good in terms of resemblance to the human voice and intonation. What makes it less quality than the human voice are some sounds, "clicks", that appear from time to time."
"The quality was quite good in the first test, while in the rest the quality deteriorated. The concatenation between units is quite noticeable."
"Quite good; the main problem are the coarticulatory artefacts that make the voice less natural."
"Quite good taking into account that it's a synthetic audio. However, it's quite noticeable that it's not a natural human voice."
"Acceptable quality. The only problem I detect, which is repeated quite often, is the lengthening/dragging of some vowels and consonants."
"The voice is correct and clear but from time to time there are strange sounds and it sounds distorted."

Source: authors.

The audio and video files generated by the experiments can be found at the following website: <http://www.salle.url.edu/portal/departaments/home-depts-DTM-projectes-info?id_projecte=67>

6. Conclusions and future lines of work

In this article, after reviewing the state of the question in the area of speech synthesis (also known as *text-to-speech conversion systems*), the status has been studied of this technology in Catalonia and, specifically, in the area of audiovisual productions. At present there are various research centres and companies working to develop and improve speech synthesis systems in Catalan. However, the penetration of these systems in the context of generating audiovisual productions is still very limited. Given this situation, the feasibility has been evaluated of implementing this technology in the world of audiovisual productions, by means of fieldwork that has consisted of several interviews, both with technical staff and potential users, as well as a set of experiments designed to study the degree of acceptance of synthesis in real examples.

From both the interviews and the experiments carried out, it can be concluded that using synthetic voices in broadcast content could become a reality in the coming years if certain aspects are improved in terms of the expressiveness of the content. Another important aspect is the number of modes that form part of the content. If the voice is accompanied by other, superimposed audios, as well as a video channel, the use of a synthetic voice is likely to be more feasible. On the other hand, in content where there is only a voice (e.g. a radio advertisement), listeners' demands regarding the quality of this voice are greater.

Using speech synthesis (not just in Catalan) as a support in order to have more automated systems capable of providing more natural and also more inclusive content is one of the chal-

Table 2. Selection of answers to the question "Do you think it's feasible to use speech synthesis to explain news on programmes generated automatically?"

"Yes, I think it's feasible and interesting."
"Yes, especially if they are short news items and breaking news, so that a semi-automatic form is more appropriate, making it possible to present the content more quickly."
"It should be more than feasible in the future."
"It wouldn't be feasible for TV news, for example, but it might be for internet content, where the quality of the content is not the priority but the content per se."
"It lacks naturalness and expressiveness, which help to make a news item more attractive. However, the intelligibility is very good and the message can be conveyed perfectly. It would be feasible."
"Yes. In spite of the lack of naturalness, which can be improved. The result is satisfactory enough."
"Yes. The small problems with the synthesis remain under the news soundtrack and don't pose a problem to understanding it. What's more, formally the narration is correct (neutral tone)."
"Yes. It's just as intelligible as the human voice."
"Yes but it depends on the sphere in which it's applied. If it's an internet platform, I think this quality is acceptable for users."
"Yes, provided the aforementioned artefacts are avoided."
"Yes, I think it's feasible but not as TTS is now. It still needs to be more natural. The voice it generates at present is too unpleasant for a narrator you have to listen to regularly."
"The understanding is perfect. If the issue with the small distortions could be improved, it would make following the new items more pleasurable."

Source: authors.

lenges that is already being tackled. In this respect, there are studies (*http://www.daisy.org/benefits-text-speech-technology-and-audio-visual-presentation*) that state that including the acoustic mode as an alternative to presenting content purely in a text format, for example, helps to increase retention capacity, therefore being a suitable way to present learning activities in environments of a more audiovisual nature. There are already companies whose business is to provide automated voice services based on textual information, such as IVO Software (*http://www.ivona.com*), Odiogo (*http://www.odiogo.com/*) and NextUp.com (*http://www.nextup.com*), with which, for example, oral information can be incorporated into a website or solutions provided to generate voices automatically based on text documents. Although solutions such as these will provide us with systems that are increasingly adapted to the user, we are still far from seeing systems that act like people and avoid any minimal sound artefact or that accentuate the differing degrees of expressiveness in the human voice. In any case, the solutions we might find in the world today still don't allow us to achieve a quality message comparable to narration by a real person in a real conversation, but we are getting closer and the new paradigms of interaction and exchange with content providers that will appear in the future will surely take into account the use of speech synthesis technology as a highly valid tool to emphasise or provide a message that is closer to a human one.

In order to enable the use of speech synthesis in audiovisual content, progress needs to be made along the following lines of investigation:

- Improve the expressiveness of generated speech, adapting suprasegmental features (pace, intonation, intensity, emphasis, etc.) to the narrative characteristics of each kind of content. This improvement can be achieved if we take advantage of the expertise in the field of audiovisual communication.
- Improve the segmental quality of the synthesis to avoid sound artefacts, as we must remember that human hearing is very sensitive to these small errors. Particularly important are errors related to phonetics and signal processing. It would therefore be advantageous to employ the know-how of phonetics experts to improve, for example, the rules of phonetic transcription, especially related to coarticulation. Regarding signal processing, there is still a long way to go in voice parameterisation and modelling to be able to modify a voice's characteristics without distorting it.
- Find new methods to generate new voices using voice transformation techniques that help to increase the number of high quality voices available in a specific language.

References

Alías, F.; Formiga, L.; Llorà, X. "Efficient and reliable perceptual weight tuning for unit-selection Text-to-Speech synthesis based on active interactive genetic algorithms: a proof-of-concept". *Speech Communication,* vol. 53 (5), p. 786-800, May-June, 2011.

Campbell, N. "Evaluation of Text and Speech Systems". *Text, Speech and Language Technology*, vol. 37 (2007), p. 29-64, Springer, Dordrecht.

Camps, J.; Bailly, G.; Martí, J. "Synthèse à partir du texte pour le catalan". *Proceedings of 19èmes Journeés d'Études sur la Parole* (1992), p. 329–333, Brussels, Belgium.

Guaus, R.; Iriondo, I. "Diphone based Unit Selection for Catalan Text-to-Speech Synthesis". *Proceedings of Workshop on Text, Speech and Dialogue* (2000), Brno, Czech Republic.

Iriondo,I.; Alías, F.; Melenchón, J.; Llorca, M. A. "Modeling and Synthesizing Emotional Speech for Catalan Text-to-Speech Synthesis". *Tutorial and Research Workshop on Affective Dialog Systems, Lecture Notes in Artificial Intelligence*, no. 3068 (2004), Springer Verlag, p. 197-208, Kloster Irsee, Germany.

Iriondo,I.; Socoró, J. C.; Alías, F. "Prosody Modelling of Spanish for Expressive Speech Synthesis". *International Conference on Acoustics, Speech and Signal Processing*, vol. IV (2007), p. 821-824, Hawaii, United States.

Llisterri, J.; Fernández, N.; Gudayol, F.; Poyatos, J. J.; Martí, J. "Testing user's acceptance of Ciber232, a text to speech system used by blind persons". *Proceedings of the ESCA Workshop on Speech and Language Technology for Disabled Persons* (1993), p. 203–206, Stockholm, Sweden.

Planet, S.; Iriondo, I.;Martínez, E.; Montero, J. A. "TRUE: an online testing platform for multimedia evaluation". Proceedings of the Second *International Workshop on Emotion: Corpora for Research on Emotion and Affect at the 6th Conference on Language Resources & Evaluation* (2008), Marrakesh, Morocco.

Rodríguez, M.A.; Escalada, J. G.; Armenta, A.; Garrido, J.M. "Nuevo módulo de análisis prosódico del conversor texto-voz multilingüe de Telefónica I+D". *Actas de las V Jornadas en Tecnología del Habla* (2008), p. 157-160.

Torrens, A. "Estudi sobre la utilització de les tecnologies de síntesi de veu en els mitjans audiovisuals de Catalunya". *Treball final de carrera* (2010). Barcelona: La Salle - Universitat Ramon Llull.

UIT-T (1996). "Recomendación P.800: Métodos de determinación subjetiva de la calidad de transmisión*". Sector de Normalización de las Telecomunicaciones de Unión Internacional de Telecomunicaciones*.
<http://www.itu.int/rec/T-REC-P.800-199608-I/es>