REVIEW OF RESEARCH ON SPEECH TECHNOLOGY: MAIN CONTRIBUTIONS FROM SPANISH RESEARCH GROUPS

SAN-SEGUNDO, Rubén*[1]
MARTÍNEZ-HINAREJOS, Carlos D.[2]
ORTEGA, Alfonso[3]

1. Grupo de Tecnología del Habla. Universidad Politécnica de Madrid.
2. Pattern Recognition and Human Language Technologies (PRHLT). Universidad Politécnica de Valencia.
3. Voice input Voice output Lab (ViVoLab). Universidad de Zaragoza.

In the last two decades, there has been an important increase in research on speech technology in Spain, mainly due to a higher level of funding from European, Spanish and local institutions and also due to a growing interest in these technologies for developing new services and applications. This paper provides a review of the main areas of speech technology addressed by research groups in Spain, their main contributions in the recent years and the main focus of interest these days. This description is classified in five main areas: audio processing including speech, speaker characterization, speech and language processing, text to speech conversion and spoken language applications. This paper also introduces the Spanish Network of Speech Technologies (RTTH. Red Temática en Tecnologías del Habla) as the research network that includes almost all the researchers working in this area, presenting some figures, its objectives and its main activities developed in the last years.

Keywords: Speech Technology; Spain; Castilian; Catalan; Basque; Galician.

1 Introduction

In the last 20 years, there has been an important expansion of research in speech technology in Spain: the number of research groups has increased from 3 to more than 20 distributed all along the Spanish geography in more than 20 different universities and research centres. Important companies like Telefónica I+D created a specific Speech Technology Division (now integrated in other divisions more oriented to multimedia applications) for developing their own products. Additionally, small spin-off companies appeared, such as Verbio and Agnitio, among others.

This expansion has been motivated by an increase in the investment from the European Commission, Spanish Ministry of Education, and local administrations, but also, by the interest of using speech as an important modality when developing Human-Computer interaction applications. At the beginning, these applications were focused on services over the telephone like information delivery or ticket reservation services. Nowadays, the main applications including speech technology are focused on developing advanced interfaces for mobile devices, automatic multimedia indexing and accessing tools, and new applications for people with special needs.

This increase in the number of research groups focused on speech technology has allowed, on the one hand, to deal with a wide range of applications involving these technologies: from audio processing for extracting speech in multimedia contents, to multimodal and multilingual spoken dialogue applications, including also speech translation, oral communication disorder detection, applications oriented to people with disabilities, etc. And, on the other hand, to face all the official languages in Spain: Castilian, Catalan, Basque and Galician, with their corresponding dialectal diversity.

This paper provides a review of the main areas of speech technology addressed by research groups in Spain, their main contributions in the recent years and the main focuses of interest these days.

The paper is organised as follows: section 2 describes the Spanish Network on Speech Technology; Section 3 and 4 are focused on Audio Processing and Speaker Characterization respectively. Section 5 describes advances on Speech and Language Processing. Section 6 describes Spoken Language Applications and, finally, section 7 summarizes the main conclusions.

2 Spanish Network on Speech Technology

Ten years ago, Prof. Antonio Rubio led a group of researchers for creating the Spanish Network of Speech Technology (RTTH: Red Temática en Tecnologías del Habla: www.rthabla.es). Nowadays, this network includes more than 250 researches from more than 20 different Spanish universities and research centres. Table 1 summarizes the most relevant figures of the main research groups involved in this network.

Table 1. Main figures of the Spanish Network on Speech Technology

Main figures (in the last 5 years)	Number
Researchers	> 250
Universities, Research Centres and Companies	> 22
European projects	> 20
Public or private projects	> 300
Research events organized	> 35
Publications in JCR journals	> 180
Papers at relevant international conferences	> 350
Patents	> 10

As it is shown, the high number of researchers reveals an important critical mass on these technologies in Spain which are responsible for a high number of EU or domestic projects, and which were responsible for important amount of research events organized during the last 5 years. This activity has generated important results like journal and conference publications, and patents.

The RTTH has three main objectives:

- The first one is to support research activities on speech technologies in order to complement the main activities developed inside target oriented research projects.
- Secondly, considering the relevance of the financial and human resources for boosting a research area, the RTTH has the target of attracting new investments (from companies and governments) and new researchers to this research field.
- Finally, the RTTH defines a collaborative framework for all Spanish research groups working on speech technologies.

The main activities performed by the RTTH during the last years have been the following:

- The RTTH has promoted every two years the "Jornadas en Tecnologías del Habla" since 2000. Previous workshops were held in Sevilla (2000), Granada (2002), Valencia (2004), Zaragoza (2006) and Bilbao (2008). The last one was organized as an international event in Vigo (November 2010) with the name FALA 2010. This workshop was a joint event including "VI Jornadas en Tecnología del Habla" and II Iberian SLTech Workshop. In these workshops, there have been relevant talks given by international experts in these technologies.
- Another important activity is to organize awards for Ph. D. students:
    - At every workshop, there are a number of awards from 3 to 6 for the best papers presented in the workshop where the first author is a Ph. D. student.
    - Additionally, every year there is an award for the best JCR paper published by a Ph. D. student as a result of his/her doctoral thesis.
- Additionally, the RTTH organizes meetings between several research groups in order to promote new research projects with the collaboration of those groups.
- Finally, RTTH collaborates with the SIG-IL (Special Interest Group on Iberian Languages) of ISCA (International Speech Communication Association).

In the next section, the paper describes the main areas of interest on which the Spanish research groups focus. This description is organized in four sections: audio processing including speech, speaker characterization, speech and language processing, and spoken language systems.

3 Audio Processing including Speech

This section integrates technologies for acoustic environment characterization, voice activity detection (VAD) and multimedia processing for audio segmentation.

3.1 Acoustic environment characterization

One important area of interest that has appeared recently is the characterization of the acoustic environment. Initially, this characterization had the target of modeling the acoustic noise in order to increase the robustness of the applications when extracting the speech from a specific acoustic environment. Nowadays, this area also includes the possibility of extracting useful information about the speaker environment: activity performed, mean of communication, existence of other speakers or sound sources, speaker location, user's current situation, and so on, which potentially enhance the description of a location and user's activities (Ma, 2006; Chu, 2008). A context denotes a location with different acoustic characteristics, such as a coffee shop, outside street, or a quiet hallway. This new application (extracting useful information) is a new research field with an important activity in Spain.

The most frequent approach has been based on recognition of specific events or sounds (Cai, 2006). Only a few systems have been proposed to model raw environment audio without pre-extracting specific events or sounds (Buera, 2007; Eronen, 2007).

When detecting and classifying acoustic events the main target is to extract a particular type of event (speech, music, coughs, …) and the time position from a mixed general acoustic material. This field of interest has grown rapidly during the last years (Ntalampiras, 2009; Portelo, 2009) proposing this detection in a wide range of possible applications such as meeting rooms, hospitals, and public places, and for audio segmentation of broadcast news. The detection task is more difficult when there is an important overlap of these acoustic events or even with speech: detecting isolated acoustic events shows a high recognition rate when considering a small number of possible events. When dealing with overlapped acoustic event, the use of complementary video information has shown its usefulness to detect audio sources. In Spain, a pioneering research work has recently been done on AED/C applied to meeting-room acoustic events in the framework of the CHIL project (http://chil.server.de) (Temko, 2009a, 2009b; Butko, 2011;). In the Albayzín 2010 Evaluation (http://fala2010.uvigo.es), there was a competition on Audio Segmentation, using a Catalan broadcast news database from the 3/24 TV channel (Albayzín, 2010). Additional research on speech and music segmentation was developed in (Gallardo-Antolín, 2010).

An important information that can be extracted from the acoustic environment is the location of the acoustic source, specially where the speaker is situated. This aspect can report valuable information for improving advanced human-computer interfaces considering speech as a main modality. Some applications of this location information are:

- To select the track a robot can follow in order to interact to the user in a better way.
- To steer a camera towards the active source, enhancing the audio stream via microphone-array beamforming for speech recognition (a challenging task in smartroom scenarios, given the severe degradation of the speech signal due to noise and room reverberation effects).
- To provide accumulated information for person identification, and to recognize location-based events (AMIDA, 2007).

Speaker localization has been investigated using computer vision systems (Fernández, 2007; Pizarro, 2009), audio source location systems (Lathoud, 2007) and mixed approaches based on audio-visual fusion (Chen, 2004; Gatica-Perez, 2007). Additionally, it is interesting to remark that microphone array speech recognition (i.e., the integration of beamformer with ASR) has been also investigated (Moore, 2003) but it still has low performance compared to close talk speech recognition. This task is even more complex in the case of several speakers talking at the same time in a meeting (Shriberg, 2001).

3.2 Voice activity detection

There are many advantages of using speech-based applications in order to improve Human Computer Interaction systems (especially over the telephone). But these applications have a poor performance when the main speaker is embedded in noisy environments (for example in bars), where many far-field speakers are speaking almost all the time. In particular, it is very common to find, in mobile phone scenarios, many situations in which the target speaker is situated in open environments surrounded by far-field interfering speech from other speakers. This factor contributes to a reduction in the speech-based application performance, producing an unsatisfactory experience for the user.

Because of this, Voice Activity Detection (VAD) is a relevant task for speech-based real-world applications considering one microphone (De la Torre, 2006), or specifically to exploit the availability of microphone array signals to improve VAD results for far field speech (Lathoud, 2007). Currently, most VADs are focused on the detection of speech acquired in noisy conditions. Some VADs take the speech-non speech decision based on statistical properties of features derived from the signal, that differ from speech to non speech periods (Górriz, 2005). Other approaches are based on models that represent speech or non-speech (De la Torre, 2006). Others VAD applications reduce the amount of signal to be processed or transmitted and define the noise parts in order to estimate the background noise features.

The most relevant databases widely used for VAD evaluation are those developed under the AURORA project (i.e. AURORA-2, -3 and -4 databases).

3.3 Multimedia processing for audio segmentation

In many applications, considering complementary modalities for extracting information can significantly improve the results. For example, the analysis of audiovisual streams for detecting and segmenting speaker activity is a helpful task for speaker diarization (González-Jiménez, 2007a, 2007b; Alba-Castro, 2008; Argones-Rúa, 2008; Argones-Rúa, 2009). Another example is face recognition that can increase the reliability of speaker identification when adapting the system to a specific speaker. On the other hand, lips tracking can be used for enhancing speech recognition, mainly in noisy environments (in the street), whenever the video resolution and quality allow extracting accurate information from the lips area (Perez-Freire, 2004). In Spain, there is a Database of TV-news acquired from TV streams, called Transcrigal-DB.

4 Speaker Characterization

This section includes speaker recognition and diarization, language recognition, emotion recognition, and voice disorder detection.

4.1 Speaker recognition and diarization

Speaker recognition consists of detecting the identity of the person who is speaking in a specific moment. Detecting this identity automatically requires two phases: enrolment and verification. During enrolment, the speaker's voice is recorded and a number of features are extracted in order to create a voice model. In the verification phase, a speech sample is compared against a previously created voice model. In the last years, there has been a predominance of a combination systems approach (fusion) for taking advantage of non-correlated information for identifying the speaker (Brummer, 2007). Non-correlated information is mainly obtained from several algorithms that analyze speaker characteristics using different pieces of information extracted from complementary levels: acoustic, phonetic, prosodic, or lexical levels.

Gaussian Mixture Model (GMM) has been widely used for modelling the speaker characteristics related to his/her identity considering information from the short-term spectral level (acoustic level). Nowadays, systems based on Support Vector Machines (SVM) have demonstrated a good performance in the task by using a discriminative approach (Campbell, 2006) in comparison to a generative approach (as in GMM). In the last five years, in order to improve the performance, there has been an increasing interest in extracting features from higher levels of information present in speech, such as pronunciation variation, linguistic content, prosody, which happened to be very useful in automatic speaker recognition (Gonzalez-Rodriguez, 2007). It is also important to remark that speech-based techniques can be combined with information coming from others modalities in order to improve the system performance (Ortega-García, 2010).

In recent years, several groups have developed speaker recognition systems based on Joint Factor Analysis (JFA) (Kenny, 2008) due to the fact that this approach allows to model several sources of variability and compensate them, increasing the performance of the systems. This trend has led to the development of systems based on i-vectors (also known as total variability factors) (Dehak, 2011), which aim at modelling the overall variability and try to compress the essential and useful information into a low-dimensional space where speakers are modelled. Thanks to these new proposed techniques the Equal Error Rate (EER) has decreased below 2% for verification tasks as the ones proposed in the Speaker Recognition Evaluations (SRE) organized by NIST, in which several Spanish groups usually take part.

An important research field that has increased its interest in the last years has been speaker diarization. This subject consists of segmenting an input audio stream into homogeneous segments according to the speaker identity, annotating speech with speaker turns. The main applications of detecting automatically speaker turns are, among others:

- To incorporate additional information for automatic audio indexing of meetings with speech from several people.
- When combined with speaker recognition and adaptation techniques, to improve speech recognition performance by adapting the acoustic modeling to a specific speaker

Speech diarization started being applied to a high quality recording scenario: broadcast news recordings with a high SNR (Ferreiros, 2000; Meignier, 2006). Nowadays, speech diarization has been applied to recordings of meetings and lectures which is a task that has shown a much higher level of error. The diarization error rate (DER) in these cases increases from 8% to 20%. This increment is due to several aspects like the existence of overlapping speakers. Very low diarization errors (around 1%) have also been achieved for two speaker telephone conversations as a supporting task for speaker verification systems (Kenny, 2010; Vaquero, 2011) based on the aforementioned Factor Analysis approach.

Speaker diarization is a combination of an unsupervised speaker segmentation (finding speaker change points in an audio stream) and speaker clustering (grouping together speech segments on the basis of speaker characteristics) which can use a bottom-up or a top-down strategy (Tranter, 2006). In early research, segmentation and clustering were performed independently in two steps. Nowadays, segmentation and clustering are done simultaneously and iteratively. One of the critical problems in speaker diarization is to define the cluster comparison measure in order to decide if two segments must be merged or not. The widely used distance measure is a modification of the Bayesian Information Criterion (Anguera, 2006; Wooters, 2007). When there are several microphones recording at the same time, it is possible to use this information for improving the performance: a successful method for joining speaker vocal tract features and speaker localization features was presented in Pardo (2007). New methods have recently appeared to compare clusters with very good performance in terms of computational complexity (Vijayasenan, 2009), and performance (Gallardo-Antolín, 2006; Anguera, 2009; Nguyen, 2009).

In the Albayzín 2010 Evaluation (http://fala2010.uvigo.es), there was a competition on Speaker Diarization, using a Catalan broadcast news database from the 3/24 TV channel 0.

4.2 Language identification

Language identification faces the problem of identifying the language used by a speaker in an audio recording. In this field, the main techniques used are very similar to those used in speaker recognition. The first research efforts showed that high-level systems performed better than acoustic systems, although there have been some improvements in acoustic systems since then. Both techniques are applied in a similar way in speaker and language fields, with high-level techniques for language recognition including phonotactic and prosodic modeling (Torre-Toledano, 2009; Caraballo, 2010). Recently, Factor Analysis approaches have also been successfully applied to the language recognition task allowing the compensation of inter-session variability (Brümmer, 2009), and outperforming the existing acoustic systems. In the Albayzín 2010 Evaluation (http://fala2010.uvigo.es), there was a competition on Language Identification, using a TV broadcast speech database (Rodríguez-Fuentes, 2010).

4.3 Emotion recognition

Using the speech signal for recognizing the speaker emotion is a new research field very interesting for the domain of human-computer interaction and affective computing dedicated mainly to three main applications:

- To improve efficiency and friendliness of human-machine interfaces.
- To allow for monitoring of mood state of individuals in demanding working tasks.
- To add information into automated medical or forensic data analysis systems (Taylor, 2005).

Recently, there has been an important research effort (Scherer, 2003), but the problem is still open (Navas, 2005; Barra-Chicote, 2010; Luengo, 2010). Although a lot of research has been done on defining good features of emotional speech signal, no widely acknowledged set of speech signal features has been defined. Nowadays, the main target is to find a feature vector for performing the classification task. Related to the emotional speech modelling and classification, it is possible to find traditional classification methods like neural network, SVM, LDA or QDA (Kwon, 2003; Fragopanagos, 2005).

4.4 Oral Communication Disorders

In the last decade, there has been in Spain an important increment of research groups working on automatic techniques for detecting oral communication disorders which includes voice pathologies, speech impairments or language impairments. Oral communication disorders are very frequent in the population: it is estimated that 20% of people suffer or have suffered from dysphonic voice. The automatic detection of oral communication disorders can be useful for medical applications both in diagnostic and therapy systems. Additionally, the information extracted from these techniques contributes to characterize the speaker, helping to improve a speaker identification system, for example. The evaluation of the voice quality by means of biometric features appears as individual problems that can be used for medical or forensic scenarios, complementing the voice characteristics. New technologies are helping to analyse videoendoscopic high and low speed sequences, allowing progress on understanding the phonatory process and establishing correlations with the parameters extracted from the acoustic record. These correlations are still not well known, and it is necessary to invest more time and effort on finding the relationship between changes at the biomechanical level with the voice register and the estimated glottal wave. One way to address this study is a multimodal approach, mixing speech and video processing techniques (Yan, 2006; Lohscheller, 2008; Zang, 2010). A relevant aspect is to define the relationship between the glottal waveform (obtained from Electroglottography) and inverse methods (by inverse filtering of voice, or by synthesis from the high or low speed videoendoscopic images).

Most state-of-the-art systems designed to detect voice or speech disorders use the speech trace and classification strategies based on statistical and probabilistic methods (Bayesian networks, HMMs, GMMs, etc.) or neural networks (MLP, RBF, SVM, etc.) (Godino-Llorente, 2004). There is also a large amount of feature extraction approaches (Godino-Llorente, 2006a, 2006b) classifying them into long-term averaged parameters (HNR, NNE, GNE, VTI, jitter, shimmer, tremor, LTAS spectrum, etc.) and short-term parameters (MFCC, LPCC, PLP, etc).

Traditionally, voice pathology detection methods have been developed and evaluated considering sustained phonation of vowels. Nowadays, the use of continuous voice is a new challenge, as well as the use of complexity measurements and biomechanical parameters estimated from the speech (Godino-Llorente, 2006b; Sáenz-Lechón, 2006; Osma-Ruiz, 2008a; Fernández-Pozo, 2009) or the use of multimodal information for generating a speech evaluation, including videoendoscopic images.

A very interesting problem in this area of research is the database generation, having the support of a medical team and the possibility of accessing a group of patients suffering from voice disorder, which is a complex problem.

5 Speech and Language Processing

This section includes Automatic Speech Recognition (ASR) and Text To Speech (TTS) conversion. ASR and TTS technologies have obtained information from speech analysis and production studies. These studies are focused on phonation and glottal level processes considering the source-filter theory by Gunnar Fant. In the last years, these studies have been applied in order to define automatic strategies for evaluating the voice quality. In relation to speech perception, the main research lines have evolved to psychophysiological studies by means of simulation strategies. This simulation has been possible thanks to the reproduction of

biophysical and perceptual phenomena in a simulation workbench. Some important references on speech production and perception models are (Gandour, 2007; Munkong, 2008; Rauschecker, 2009; Gómez, 2009a, 2009b).

5.1 Automatic Speech Recognition

In the state-of-the-art, all speech recognizers developed so far are based on two sources of knowledge: phone acoustic characterization and language structure. Their objective is to reduce the word error rate (WER) of the speech recognition system, that is, the number of misrecognized words, to the minimum.

In almost all current speech recognition systems, the acoustic modelling is based on Hidden Markov Models (HMMs). For each allophone (a characteristic pronunciation of a phoneme), one HMM model is calculated as a result of a training process carried out using a speech database. A speech database consists of several hours of transcribed speech (composed of files with speech and text combined, where it is possible to relate the speech signal to the words pronounced by the person). In the 1970-1980s some authors such as Baum (1972), Jelinek (1976), and Rabiner (1988), contributed in a decisive way to speech recognition research by establishing the basis of the theory of the Hidden Markov Models (HMM), which have survived to our time. But it is also true that this modeling is not enough and it is still necessary to invest much effort in order to reach a performance like that of human beings in the same conditions.

Given that HMMs are statistical models, it is possible to modify and adapt the model parameters to reduce the word error rate (WER), by adapting the models to special acoustic characteristics: high level of noise (Buera, 2007; Jinyu, 2007), or speaker variability (Leggetter, 1995; Lee, 1998). These basic techniques have been expanded in more recent works like Miguel (2008). It is possible to adapt the model to a new task by discriminative training by taking errors into account for correcting the models (Jiang, 2006). In the field of very large vocabulary systems, there are many recent innovations (Aubert, 2002; Livescu, 2007) which have decreased the WER while achieving reasonable computing times.

Statistical performance depends strongly on the amount of data used to train the models. Database acquisition is a very costly process because it requires linguistic experts for manually transcribing the speech pronounced by different speakers. Because of this, only important companies or important research centres with a large experience in this technology can offer speech recognition systems with the highest warranty of having enough robustness and flexibility. In the speech community, there are two main associations that sell speech databases for research and development: LDC (Linguistic Data Consortium: http://www.ldc.upenn.edu/) and ELRA (European Language Resources Association: http://www.elra.info/). The main speech databases collected for Spanish, Catalan, Galician, and Basque have been developed and generated in Spanish centres. Some examples are Albayzin (Moreno, 1993) and the database collected in the SpeechDat project (van den Heuvel H et al, 2001).

Research groups in Galicia, Catalonia, and the Basque Country have developed ASR systems in Catalan, Galician and Basque with characteristics comparable to those in Castilian.

The second source of knowledge included in a speech recognizer is the language modeling (Nadas, 1984). This model complements the acoustic knowledge with information about the most probable sequences of words. The language modelling (LM) task is stated as the problem of designing appropriate models that approximate the probability of a given text. Therefore, given a sentence or text made up of several words, the main target of LM is to model the probability of each word, given the model. Generally speaking, there are two main

approaches for language modelling: statistical- and grammar-based language models. In these approaches, there are several models for approximating the actual language probability distribution. For instance, hierarchical models use context-free grammars to capture long term dependencies (Benedí, 2005). However, one of the most widespread models is the n-gram model (Rosenfeld, 1994; Goodman, 2001), which obtains surprisingly good performance although it only captures short-term dependencies.

A very attractive research area that has increased its interest has been Remote Speech Recognition (RSR). This increment has been due to the fast development of wireless networks and mobile devices connected to them. The main problem of these devices is their size and weight. The size opens new possibilities to apply speech recognition in the interface, but the weight limits the computation power of these devices. RSR tries to overcome these constraints by moving the most complex computational tasks of speech recognition to a remote server (Peinado, 2005; Gómez, 2006, 2007, 2009; Carmona, 2010). There are two possibilities for the implementation of an RSR system: Network Speech Recognition (NSR), where the whole recognition system resides in the network, and the speech signal is sent through the network, or Distributed Speech Recognition (DSR), where the client includes a local front-end that processes the speech signal in order to obtain the specific features used by the remote server (back-end) to perform recognition. NSR systems do not require modifications on the client terminal, although speech needs to be coded in order to reduce the traffic. This also involves information loss that may affect speech recognition performance. The ETSI STQ-Aurora working group has been working on defining a standard front-end to facilitate its integration in commercial mobile devices.

Without any doubt, the main problem of the ASRs is to deal with noisy speech. In these circumstances, the ASR performance degrades considerably. In order to avoid this degradation, there are two main research tendencies. The first one consists of enhancing the speech to improve its perceptual quality by reducing the acoustic noise. In this line, traditional techniques such as Spectral Subtraction and Wiener filtering are being widely used (Gallardo-Antolín, 2002). In recent years, different techniques have been proposed for single-channel and multiple-channel speech enhancement (a review can be found in (Krishnamoorthy, 2009)). Other approaches, more suitable for dealing with multi-speaker environments, are based on the enhancement of LP residuals or Computational Auditory Scene Analysis (CASA). For the case of multi-channel methods, speech enhancement is provided by exploiting the spatial diversity produced by the different locations of desired and undesired sound sources in space (Maganti, 2007).

Secondly, another strategy is to extract more robust features from speech in order to reduce the degradation produced by the noise. Some approaches have been proposed based on feature normalization or filtering of the temporal trajectories of the acoustic parameters (Nadeu, 1997, 2001; Vicente-Peña, 2006). A review of some of these methods can be found in Peinado (2006).

5.2 Text to speech conversion

Similar to ASR, the other important research area in speech technology during the last ten years has been Text To Speech (TTS) conversion. In this area, there are mainly two approaches: unit selection and statistical parametric speech synthesis.

- The unit selection has been the main technique during the last twenty years. This technique reaches quite natural speech by concatenating acoustic and prosodic units selected from a large corpus (Hunt, 1996; Raux, 2003; Navas, 2006; Escudero, 2007)

containing hundreds of realizations of each phoneme in different contexts, so that the amount of signal processing required after the concatenation is minimal. This approach has several problems: the first one is that the system quality depends significantly on the amount of speech data available (the number of speech chunks to concatenate). Secondly, this strategy has a reduced flexibility: the synthetic speech is strongly conditioned by the content of the corpus in terms of style, speaker, dialect etc. When some of these aspects must be changed, it is necessary to record a new database. In order to avoid these problems, voice transformation techniques have been proposed (Stylianou, 2009), but they have not reached the level of performance obtained with statistical TTS.

- The second technique is based on Hidden Markov Models (HMMs) (Zen, 2009). This approach has become very popular due to the high degree of flexibility that results from the statistical parametric representation of the voice. Such systems can generate speech adapted to different styles, speakers (Erro, 2010), or dialects (Yamagishi, 2009). However, the quality of the synthetic speech is degraded by the limitations of the parameterization, the modelling capabilities of HMMs, and oversmoothing.

Using statistical TTS has permitted to adapt a TTS to any conditions in an easier way. Special interest has appeared in adapting the TTS for generating different emotions (Barra-Chicote, 2010; Erro, 2010).

Research groups in Catalonia, Galicia, and the Basque Country have developed TTS for Catalan (Bonafonte, 2008), Galician (García-Mateo, 1998; González-González, 2004), and Basque (Navas, 2002) with comparable characteristics to Castilian ones.

6 Spoken Language Applications

This section includes spoken language understanding and translation, spoken dialogue systems, voice-activated question answering (QA), and applications for people with special needs.

6.1 Spoken Language Understanding

This process consists of extracting semantic information or "*meaning*" (related to the specific application domain) from the speech recognizer output (sequence of words). Semantic information is represented by means of semantics concepts. A semantic concept consists of an identifier and a value (sequence of words that generated the concept). For example: we could have a concept TURN while the value is "to the right". We could classify the language understanding techniques in two types: statistical (or data-driven) and rule-based techniques.

- ***Data-driven approaches.*** Many data-driven systems depend on statistical models to derive the corresponding semantic representation from an input utterance. A simple but effective semantic decoding model is the Hidden Markov Model (HMM), which was adopted in the AT&T's CHRONUS (Pieraccini, 1993). Wang (2001) proposed a semi-automatic grammar learning methodology by taking advantage of multiple information sources, such as automatically generated template grammar from semantic schema, the semantically annotated corpus, syntactic constraints, and grammar library. In order to infer a good quality grammar, the grammar learning approaches often require a large amount of annotated data or linguistic experts.
- ***Rule-based techniques.*** In this case, the relations between semantic concepts and word sequences are defined manually by an expert. These approaches have been based

on grammar-based parsers interleaving syntax and semantics (Seneff, 1992; Dowding, 1993), or purely semantic (Ward, 1994; Wang, 1999). The rule-based techniques can also be classified into two types: top-down and bottom-up strategies. In the first case, the rules are conceived in such a way to obtain the semantic concepts from a global analysis of the whole sentence. In the bottom-up strategy, the semantic analysis is performed starting from each word individually and extending the analysis to neighborhood context words. This extension is done to find specific combinations of words that generate a semantic concept.

An emerging trend of spoken language understanding is to combine the rule-based and data-driven methods in order to make use of their advantages (Wang, 2002).

6.2 Spoken language translation

Related to spoken language translation, statistical approaches have achieved performance levels comparable to those achieved by knowledge-based Machine Translation (MT) algorithms, which have been around for more than half a century. The best performing translation systems are based on various types of statistical approaches, including example-based methods, finite-state transducers and other data-driven approaches. The progress achieved over the last 10 years is due to several factors such as efficient algorithms for training, context dependent models, efficient algorithms for generation, more powerful computers and bigger parallel corpora, and automatic error measurements.

Specifically in the European Community, where the language diversity still represents an important drawback for the integration process, a large amount of resources has been invested in R&D in this technology. As a representative example of such effort, the following projects can be mentioned: C-Star, Eutrans, Verbmobil, LC-Star, Nespole!, Fame, TC-Star, SMART, EURO-MATRIX and PHAUST (Och, 2003; Chiang, 2007; Koehn, 2007; Waibel, 2008; PHAUST, 2010). The consortium has participated in several of them. In Spain the main research project focused on language translation between the official languages in Spanish is AVIVAVOZ (Mariño, 2006). However, no matter the recent progress, MT technology is still far from achieving satisfactory performance and quality levels (Casacuberta, 2004; Crego, 2006; Mariño, 2006; Gispert, 2008; Costa-Jussà, 2009a, 2009b). Other groups have focused MT algorithms on translating speech into sign language (San-Segundo, 2008), allowing deaf people to access to spoken language contents.

There is a large parallel and monolingual corpus for developing statistical MT systems in several languages as Catalan, Spanish, English, Arabic, Chinese, German, Italian, French, and other European languages. In particular, the Catalan-Spanish corpus includes 10 years of the paper edition of a local newspaper containing 100 millions of words.

6.3 Spoken Dialogue Systems

Design, implementation and evaluation of Spoken Dialogue Systems are complex tasks which can involve many different areas described previously (González-Ferreras, 2009): voice signal processing, speech synthesis and recognition, speaker and context characterization, language modelling, and general spoken act planning.

Similar to language modeling and understanding, the dialogue modeling techniques can be divided into two main types: statistical (or data-based) and knowledge-based modelling. In any case, the dialogue modelling techniques rely on the use of dialogue meaningful units which

are usually coded as Dialogue Acts (DA). The definition of the set of DA labels is usually related to the specific task the dialogue system must cope with.

The statistical modelling strategies consist of defining a statistical model (Partially Observable Markov Decision Processes - POMDP or Bayesian Belief Networks) and training the model parameters using an important amount of data (Griol, 2008; Sanchís, 2008; Fernández, 2009). In this case, it is not necessary to be an expert to design the dialogue but a lot of data (transcribed dialogues in similar conditions) are needed in order to train the model. POMDP are regarded as the most powerful state-of-the-art models for dialogue systems, but they need huge amounts of data to train the parameters of the model (which include transition, emission, and reward function parameters). Bayesian Belief Networks are a special case because although it is a statistical model, this paradigm can also be used as a framework for expert knowledge representation if there are not enough data for training.

In the case of knowledge-based dialogues, it is necessary to define a dialogue model to be used as implementing architecture for representing the expert knowledge. Within the speech community, we can find several dialogue models based on expert knowledge. The most important ones are the following:

- *Finite State Machine Model.* The dialogue model is based on a finite state machine: each state is associated with a different dialogue turn (San-Segundo, 2005). The dialogue flow is specified by the state order in the finite state machine. The main advantage of this model is its simplicity: when we define the state sequence, the dialogue flow is fixed. On the other hand, this model presents a high degree of rigidity and does not allow mixed-initiative dialogue managers to be developed.
- *Frame-based model.* In this case, the dialogue model is based on frames (D'Haro, 2005). For every goal in the dialogue, we define a frame associated to it. A dialogue goal is part of the functionality provided by the system. This model is more complex, but is more flexible and allows mixed initiative dialogues by both the user and the system. The user can say the commands in any order and the same order-independent behaviour happens for the command items specification.
- *Hierarchical* model. The hierarchical model is similar to the frame model. The difference is that, in this case, a new representation level between the goal and their items is introduced. In this case, a hierarchical structure is used to represent the main goals/actions of the interface, their sub-goals/sub-actions and the items associated with each sub-action.

In Spain, in the last 5 years the most relevant projects related to spoken dialogue systems are EDECAN (www.edecan.es) and SD-TEAM (www.sd-team.es).

6.4 Voice-activated Question Answering Applications

Another important area of interest is developing Voice-activated Question Answering (QA) applications. This interest is due to the need of providing mobile devices with more attractive functionalities. There is also an increasing interest in improving access to information systems, particularly in the way this information can be accessed, i.e., voice interfaces, mobile devices, etc. But today, only research prototypes have been developed (Harabagiu, 2002; Hori, 2003; Sanchís, 2006; González, 2008; Turmo, 2009; Sibel, 2009a, 2009b). Progress in QA depends strongly on advances on ASR technologies applied to open vocabulary domains. There are many examples of these kinds of possible applications: car navigation, tourist or cultural information search, etc. E-learning can also be a suitable field for applying this kind of technology since only short answers are frequently required for learning purposes (names or

dates in history, names in medicine, definitions or formulae in technical studies, etc.). A lot of this information is not structured and is not available in a database, but it is provided as row texts. In Europe, an important project is QALL-ME (http://qallme.itc.it/), which is focused on developing QA systems in mobile applications for information search and multimodal output (text, images, and videos). In Spain, the main domestic project dealing with these technologies is BUCEADOR (Moreno, 2010).

6.5 Multimodal applications

Speech recognition can be used in applications where speech is only one of the forms of communicating with the system. For example, speech input can be combined with touchscreen input or with handwritten strokes. The combination of speech with the other modalities can improve the performance of the whole system, since the other sources restrict the possible hypotheses that can be recognized by the system.

One recent example is the combination of handwritten text recognition with speech recognition, which benefits from the fact that both tasks use similar models (HMM and n-grams) and processes (Viterbi decoding) to obtain the results. In this case, previous recognition of text or speech can be used to restrict the decoding process in the other modality, which in general will result in a gain of performance of the whole system (Toselli, 2010).

6.6 Applications for people with special needs

A very interesting field of applications of speech technologies are those oriented to help people with special needs: voice impaired, handicapped, elderly, blind, or hearing impaired. In this field, it is worthy citing the VIVOCA project (http://www.shef.ac.uk/cast/projects/vivoca). This projects aims at developing a portable (eventually body-worn) speech-in/speech-out communication aid for people with disordered or unintelligible speech, initially concentrating on people with moderate to severe dysarthria (people who have difficulty in controlling and co-ordinating the muscles used in speech).

A large number of people have speech and language impairments beyond the phonological level, reaching the lexical level. Two different approaches are usually followed for detecting voice or speech abnormalities, the first one is based on acoustic features (Gómez, 2005) and the second one on statistical methods (Godino-Llorente, 2004; Arias-Londoño, 2010, 2011). Phonological disorders can be compensated using the traditional systems of speaker adaptation or adapting the characteristics of the vocal tract. However, when the distinctive features of the phonemes are modified, it is necessary to detect these disorders in order to define alternative pronunciations or lexical models. So, new types of modelling which overcome the limitations of acoustic model adaptation can be developed using distinctive phonetic features (sonorant, nasal, etc.). Data-driven lexical adaptation techniques have been proposed (Saz, 2009a) showing very interesting results for adapting the special lexical variability that these users present; with the interaction between acoustic and lexical adaptation frameworks as a very interesting outcome of this proposed approach (Saz, 2009b). Nevertheless, further work is necessary in order to obtain results that could allow the development of real-work application in this field.

In the field of computer-assisted tools, the use of Speech Technology can be very helpful to confirm an initial diagnosis providing an objective determination of the impairment in a non-invasive way. On the other hand, Computer-Aided Speech and Language Therapy (CASLT) tools (Saz, 2009c) are oriented to the improvement of speech quality in speakers with speech

impairments helping practitioners and complementing traditional speech therapy tools. Public institutions have been very interested in the development of this kind of tools, as can be seen in several 5th Framework Program projects of the European Union like OLP (Oester, 2002), SPECO (Vicsi, 1999) or ISAEUS (García-Gómez, 1999). Vocaliza, Preligua (Saz, 2009d) or WPCVox (Godino-Llorente, 2006) are just a few examples of these important therapy and diagnostic aid tools.

A major issue when developing new techniques or application for people with disordered speech is the availability of appropriate databases. Important efforts have been made to collect disordered speech corpora in Spanish. Among the most important databases we must mention the HACRO (Navarro-Mesa, 2005) and the Alborada (Saz, 2008) corpora.

Other set of applications is the audio description systems that could open the audiovisual world to blind people and to allow them to use speech recognition and synthesis technologies as a natural interface with the world. Finally, it's worth noting that audio subtitling and sign language translation (San-Segundo, 2008) are essential for hearing impaired people.

7 Conclusions

This paper has introduced the Spanish Network of Speech Technologies as the research network that includes almost all the researchers working in this area, signalling some figures, objectives and main activities developed in the last years. This paper has also described a review of the main areas of speech technology addressed by research groups in Spain, their main contributions in the recent years and the main focuses of interest these days. This review has been organized in four main areas: audio processing including speech, speaker characterization, speech and language processing and spoken language application. Each area includes several aspects or applications of speech technologies.

Acknowledgements

The authors want to thank all the contributions from their colleagues working in all the research groups included in the Spanish Network of Speech Technology.

Many of the research projects described in this paper have been supported by the European Commission in several Framework Programmes and by the Spanish Science and Education Ministry under many grants. The Spanish Network on Speech Technologies has been funded with the followings grants: TIC2000-2981-E, TIC 2002-11271-E, TEC2005-24712-E, TEC2006-28101-E, and TEC2009-06876-E.

REFERENCES

Alba-Castro JL, González-Jiménez D, Argones-Rúa E, González-Agulla E, Otero-Muras E, García-Mateo C. Pose-corrected face processing on video sequences for webcam-based remote biometric authentication. Journal of Electronic Imaging. 2008;18:11004-8.

ALBAYZÍN evaluation 2010; 2010 [accessed Jun 2010]. Available from: http://fala2010.uvigo.es/index.php?option=com_content&view=article&id=57&Itemid=65&lang=es.

AMIDA Project. State of the art overview: localization and tracking of multiple interlocutors with multiple sensors. Technical paper AMIDA Consortium; 2007 [accessed Mar 2010]. Available from: http://www.amiproject.org/ami-scientific-portal/documentation/annual-reports/pdf/SOTA-Localization-and-Tracking-Jan2007.pdf.

Anguera X, Aguilo M, Wooters C, Nadeu C, Hernando J. Hybrid speech/non-speech detector applied to speaker diarization of meetings. In: IEEE, editor. Proc. of IEEE Odyssey: The Speaker and Language Recognition Workshop. 2006 Jun 28-30; San Juan, Puerto Rico. p. 1-6.

Anguera X, Wooters C, Pardo JM. Robust speaker diarization for meetings: ICSI RT06s meetings evaluation system. Lecture Notes in Computer Science. 2006;4299:346-58.

Argones-Rúa E, Alba-Castro JL, García-Mateo C. On the use of quality measures in face and speaker identity verification based on video and audio streams. IET Signal Processing. 2009;3(4):301-9.

Argones-Rúa E, Bredin H, Mateo CG, Chollet G, Jiménez DG. Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden Markov models. Journal of Pattern Analysis and Applications. 2009;1;271-84.

Arias-Londoño JD, Godino-Llorente JI, Sáenz-Lechón N, Osma-Ruiz V, Castellanos-Domínguez G. An improved method for voice pathology detection by means of a HMM-based feature space transformation. Pattern recognition. 2010;23(9):3100-12.

Arias-Londoño JD, Godino-Llorente JI, Sáenz-Lechón N, Osma-Ruiz V, Castellanos-Domínguez G. Automatic detection of pathological voices using complexity measurements, noise parameters and mel-cepstral coefficients. IEEE Transactions on Biomedical Engineering. 2011;58 (2):370-7.

Aubert X. An overview of decoding techniques for large vocabulary continuous speech recognition. Computer Speech & Language. 2002;16:89-114.

Barra-Chicote R, Fernández F, Lutfi S, Lucas-Cuesta JM, Macias-Guarasa J, Montero JM, San-Segundo R, Pardo JM. Acoustic emotion recognition using dynamic bayesian networks and multi-space distributions. In: ISCA, editor. Proc. of Interspeech. 2009 Sep 6-10; Brighton, UK. p. 336-9.

Barra-Chicote R, Yamagishi J, King S, Montero JM, Macias-Guarasa J. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. Speech Communication. 2010;52(5):394-404.

Baum LE. An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities. 1972(3):1-8.

Benedí JM, Sánchez JA. Estimation of stochastic context-free grammars and their use as language models. Computer Speech and Language. 2005;19(3):249-74.

Bonafonte A, Moreno A, Adell J, Agüero PD, Banos E, Erro D, Esquerra I, Perez J, Polyakova T. The UPC TTS system description for the 2008 Blizzard challenge. Blizzard. 2008 Sep 22-26; Brisbane, Australia. p. 1-6.

Brummer N, Burget L, Cernocky J, Glembek O, Grezl F, Karafiat M, van Leewen DD, Matejka P, Scwartz P, Strasheim A. Fusion of heterogeneous speaker recognition systems in the STBU submission for the NSIT speaker recognition evaluation 2006. IEEE Transactions on Acoustics, Speech and Signal Processing. 2007;15(7):2072-84.

Brümmer N, Strasheim A, Hubeika V, Matějka P, Burget L, Glembek O. Discriminative acoustic language recognition via channel-compensated GMM statistics. In: ISCA, editor. Proc. of Interspeech. 2009 Sep 6-10; Brighton, UK. p. 2187-90.

Buera L, Miguel A, Lleida E, Saz O, Ortega A. Robust speech recognition with on-line unsupervised acoustic feature compensation. In: IEEE, editor. Proc. of IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU). 2007 Dec 9-13; Kyoto, Japan. p. 105-10.

Buera L, Miguel A, Lleida E, Ortega A, Saz, O. Cross-Probability Model based on GMM for Feature Vector Normalization in Car Environments. Biennial on DSP for in-Vehicle and Mobile Systems. 2007 Jun 1-6; Istanbul, Turkey. p. 1-6.

Butko T, Canton-Ferrer C, Segura C, Giró X, Nadeu C, Hernando J, Casas JR. Acoustic event detection based on feature-level fusion of audio and video modalities. EURASIP Journal on Advances in Signal Processing. 2011;2011:11 pages. Article ID 485738. DOI:10.1155/2011/485738.

Cai R, Lu L, Hanjalic A, Zhang H, Cai L-H. A flexible framework for key audio effects detection and auditory context inference. IEEE Trans on Audio, Speech and Language Processing. 2006;14(3):1026-39.

Campbell WM, Campbell JP, Reynolds DA, Singer E, Torres-Carrasquillo PA. Support vector machines for speaker and language recognition. Computer Speech and Language. 2006;20:210-29.

Caraballo MA, D'Haro LF, Cordoba R, San-Segundo R, Pardo JM. A discriminative text categorization technique for language identification built into a PPRLM System. In: Proc. of FALA; 2010 Nov 10-12; Vigo, Spain. p. 193-6.

Carmona JL, Peinado AM, Perez-Cordoba JL, Gomez AM. MMSE-based packet loss concealment for CELP-coded speech recognition. IEEE Tr. Audio Speech Lang. Proc., 2010;18(6);1341-53.

Casacuberta F, Vidal E. Machine translation with inferred stochastic finite-state transducers. Computational Linguistics. 2004;30(2):205-25.

Chen Y, Rui Y. Real-time speaker tracking using particle filter sensor fusion. Proc. of the IEEE, 2004;92(3): 485-94.

Chiang D. Hierarchical phrase-based translation. Computational Linguistics. 2007;33(2):201-28.

Chu S. Unstructured audio classification for environment recognition. Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence; 2008 Jul 13-17; Chicago, Illinois, USA. p.1845-6.

Costa-Jussà M, Fonollosa JAR. An Ngram-based reordering model. Computer Speech and Language. 2009;23(3):362-75.

Costa-Jussà M, Fonollosa JAR. State-of-the-art word reordering approaches in statistical machine translation. IEICE Transactions on Information and Systems. 2009;92(11):2179-85.

Crego JM, Marino JB. Improving SMT by coupling reordering and decoding. Machine Translation. 2006;20(3):199-215.

Dehak N, Kenny P, Dehak R, Dumouchel P, Ouellet P. Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech and Language Processing. 2011;19(4):788-98.

De la Torre Á, Ramírez J, Benítez C, Segura JC, García L, Rubio AJ. Noise robust model-based voice activity detection. 9th International Conference on Spoken Language Processing, Interspeech. 2006 Sep 17-21 Pittsburgh, USA. p.1954-7.

D'Haro LF, de Córdoba R, Ferreiros J, Hamerich SW, Schless V, Kladis B, Schubert V, Kocsis O, Igel S, Pardo JM. An advanced platform to speed up the design of multilingual dialog applications for multiple modalities. Speech Communication. 2005;48(8):863-87.

Dowding J, Gawron J, Appelt D, Bear J, Cherny L, Moore R, Moran D. 1993. GEMINI: a natural language system for spoken language understanding. In: Proc. of ACL. 1993 Jun 22-26; Columbus, Ohio, USA. p. 54-61.

Eronen A, Peltonen V, Tuomi J, Klapuri A, Fagerlund S, Sorsa T, Lorho G, Huopaniemi J. Audio-based context recognition. IEEE Trans on Speech and Audio Processing. 2006;14(1):321-9.

Erro D, Moreno A, Bonafonte A. Voice conversion based on weighted frequency warping. IEEE Trans. Audio, Speech and Lang. Proc. 2010;18(5):922-31.

Erro D, Navas E, Hernaez I, Saratxaga I. Emotion Conversion based on Prosodic Unit Selection. IEEE Trans. Audio, Speech and Lang. Proc. 2010;18(5):974-83.

Escudero D, Cardeñoso V. Applying data mining techniques to corpus based prosodic modeling. Speech Communication. 2007;49(3):213-29.

Fant, G. Acoustic theory of speech production. The Hague, Netherlands: Mouton, 2nd edition, 1970.

Fernández I, Mazo M, Lázaro JL, Pizarro D, Santiso E, Martín P, Losada C. Guidance of a mobile robot using an array of static cameras located in the environment. Autonomous Robot. 2007;23(4):305-24.

Fernández R, Ferreiros J, Córdoba R, Montero JM, San Segundo R, Pardo JM. A bayesian networks approach for dialog modeling: the fusion BN. In: IEEE, editor. Proc. of the ICASSP 2009. IEEE International Conference on Acoustics, Speech, and Signal Processing. 2009 Apr 19-24; Taipei, Taiwan. p. 4789-92.

Fernández-Pozo R, Murillo JLB, Gómez LH, Gonzalo EL, Ramírez JA, Toledano DT. Assessment of severe apnoea through voice analysis, automatic speech, and speaker recognition techniques. EURASIP Journal on Advances in Signal Processing - Special issue on recent advances in biometric systems: a signal processing perspective. 2009;2009: 11 pages. Article ID 982531. DOI:10.1155/2009/982531.

Ferreiros J, Ellis D. Using acoustic condition clustering to improve acoustic change detection on broadcast news. Proc. of ISCA ICSLP 2000. 2000 Oct 16-20; Beijing, China; p. 568-571.

Fragopanagos N, Taylor JG. Emotion recognition in human-computer interaction. Neural Networks. 2005;18(4):389-405.

Gallardo-Antolín A. Reconocimiento de habla robusto frente a condiciones de ruido aditivo y convolutivo [PhD Thesis]. Madrid, Spain: Universidad Politécnica de Madrid; 2002.

Gallardo-Antolín A, Anguera X, Wooters C. Multi-stream speaker diarization systems for the meetings domain. In: ISCA, editor. Proc. ICSLP 06. 17-21 Sep; Pittsburg, USA. 2006. p. 2186-9.

Gallardo-Antolín A, Montero JM. Histogram equalization-based features for speech, music, and song discrimination. IEEE Signal Processing letters. 2010;17(7):659-62.

Gandour J, Tong Y, Talavage T, Wong D, Dzemidzic M, Xu Y, Li, X., Lowe M. Neural basis of first and second language processing of sentence-level linguistic prosody. Human Brain Mapping. 2007;28:94-108.

García-Gómez R, López-Barquilla R, Puertas-Tera J-I, Parera-Bermúdez J, Haton M-C, Haton J-P, Alinat P, Moreno S, Hess W, Sanchez-Raya M-A, Martínez-Gual E-A, Navas-Chabeli-Daza JL, Antoine C, Durel M-M, Maurin G, Hohmann S. Speech training for deaf and hearing impaired people: ISAEUS consortium. In: ISCA, editor. Proc. Eurospeech-Interspeech; 1999 Sep 5-9. Budapest, Hungary. p. 1067-70.

García-Mateo C, González-González M. An overview of the existing language resources for Galician. In: ISCA, editor. LREC Workshop: Language Resources for European Minorities Languages. 1998 May 28-30; Granada. Spain. p. 1-6.

Gatica-Perez D, Lathoud G, Odobez J-M, McCowan I. Audio-visual probabilistic tracking of multiple speakers in meetings. IEEE Transactions on Audio, Speech, and Language Processing. 2007;15:601-16.

Gispert A, Mariño JB. On the impact of morphology in English to Spanish statistical MT. Speech Communication. 2008;50(11-12):1034-46.

Godino-Llorente J-I, Gómez-Vilda P. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. IEEE Transactions on Biomedical Engineering. 2004;51(2):380-4.

Godino-Llorente JI, Gómez-Vilda P. Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters. IEEE Trans. on Biomedical Eng. 2006;53(3):1943-53.

Godino-Llorente JI, Sáenz-Lechón N, Osma-Ruiz V, Gómez-Vilda P, Aguilera S. An integrated tool for the evaluation of voice disorders. Medical Engineering and Physics. 2006;28(3):276-89.

Gómez A, Peinado AM, Sánchez V, Carmona JL. A robust scheme for distributed speech recognition over loss-prone packet channels. Speech Comm. 2009;51(4):390-400.

Gómez A, Peinado AM, Sánchez V, Rubio AJ. On the Ramsey class of interleavers for robust speech recognition in burst-like packet loss. IEEE Tr. Audio Speech Lang. Process. 2007;15(4):1496-9.

Gomez A, Peinado AM, Sanchez V, Rubio AJ. Combining media-specific FEC and error concealment for robust distributed speech recognition over loss-prone packet channels. IEEE Tr. Multimedia. 2006;8(6):1228-1238.

Gómez P, Fernández-Baíllo R, Rodellar V, Nieto V, Álvarez A, Mazaira LM, Martínez R, Godino JI. Glottal source biometrical signature for voice pathology detection. Speech Communication. 2009a ;51:759-81.

Gómez P, Ferrández JM, Rodellar V, Fernández R. Time-frequency Representations in Speech Perception. Neurocomputing. 2009b;72:820-30.

Gómez P, Lázaro C, Fernández R, Nieto A, Godino JI, Martínez R, Díaz F, Alvarez A, Murphy K, Nieto V, Rodellar V, Fernández F-J. Using biomechanical parameter estimates in voice pathology detection. In: Proc of 4th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MABEVA05). 2005 Oct 29-31; Florence, Italy. p. 29-31.

Gonzalez C, Cardeñoso V, Sanchis E. Experiments in speech driven question answering. 2008. In: proc of the IEEE Workshop on Spoken Language Technology. 2008 Dec 15-19. Goa, India. p. 85-8.

González-Ferreras C. Estrategias para el acceso a contenidos web mediante habla [Ph.D Thesis]. Valladolid, Spain: Universidad de Valladolid; 2009.

González-González M. A síntese de voz en lingua galega: o proxecto Cotovía. Revista Galega do Ensino. 2004;44:199-215.

González-Jiménez D, Alba-Castro JL. Shape-driven gabor jets for face description and authentication. IEEE Transactions on Information Forensics and Security. 2007;2(4):769-80.

González-Jiménez D, Alba-Castro JL. Towards pose invariant 2D face recognition through point distribution models and facial symmetry. IEEE Transactions on Information Forensics and Security. 2007;2(3):413-29.

Gonzalez-Rodriguez J, Rose P, Ramos D, Toledano DT, Ortega-Garcia J. Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. IEEE Transactions on Audio, Speech and Language Processing. 2007;15(7):2104-15.

Goodman J. A bit of progress in language modelling. Computer Speech and Language, 2001;15(4):403-34.

Górriz JM, Ramírez J, Segura JC, Hornillo S. Voice activity detection using higher order statistics. In: IEEE, editor. Proc. of IWANN 2005 8th International Work-Conference on Artificial Neural Networks. 2005 Jun 8-10; Barcelona. Spain. p. 837-44.

Griol D, Hurtado LF, Segarra E, Sanchis E. A statistical approach to spoken dialog systems design and evaluation. Speech Communication. 2008;22:666-82.

Harabagiu S, Moldovan D, Picone J. Open-domain voice-activated question answering. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002); 2002 Aug 24-Sep 1; Taipei, Taiwan. p. 321-7.

Hori C, Hori T, Isozaki H, Maeda E, Katagiri S, Furui S. Study on spoken interactive open domain question answering. In: Proc. of IEEE & ISCA Workshop on Spontaneous Speech Processing and Recognition (SSPR). 2003 April 13-16; Tokyo, Japan. p. 111-4.

Hunt A, Black A. Unit selection in a concatenative speech synthesis system using a large speech database. In: IEEE, editor. Proc. of International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96. 1996 May 15-19. Toulouse. France. p. 373-6.

Jelinek, F. Continuous Speech Recognition by Statistical Methods. Proceedings of the IEEE. 1972;64:532-556.

Jiang H, Li X, Liu C. Large margin hidden markov models for speech recognition. IEEE Transactions on Audio, Speech and Language Processing. 2006;14(5):1584-95.

Kenny P, Ouellet P, Dehak N, Gupta V, Dumouchel P. A Study of inter-speaker variability in speaker verification. IEEE Transactions on Audio, Speech and Language Processing. 2008;16(5):980-988.

Kenny, P., Reynolds, D. and Castaldo, F. Diarization of Telephone Conversations using Factor Analysis. IEEE Journal of Selected Topics in Signal Processing. 2010;4(6):1059-70.

Koehn P. Statistical machine translation. Cambridge, UK: Cambridge University Press; 2007.

Krishnamoorthy P, Mahadeva Prasanna SR. Temporal and spectral processing methods for processing of degraded speech: a review. IETE Tech Rev. 2009;26:137-48.

Kwon O., Chan K., Hao J., Lee T. Emotion Recognition by Speech Signals. In: ISCA, editor. Proc. of ISCA Eurospeech; 2003 Sep 1-4; Geneva, Switzerland. p. 125-8.

Lathoud G, Odobez J-M. Short-term spatio-temporal clustering applied to multiple moving speakers. IEEE Transactions on Audio, Speech & Language Processing 2007;15(5):1696-710.

Lee L, Rose R. A frequency warping approach to speaker normalization. IEEE Transactions on Speech and Audio Processing. 1998;1(6):49.

Leggetter CJ, Woodland PC. Maximum likelihood linear regression for speaker adaptation of the parameters of continuous density hidden Markov models. Computer Speech and Language. 1995;9:171-185.

Li J, Deng L, Yu D, Gong Y, Acero A. High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor series. In: IEEE, editor. Proceedings of IEEE Workshop on ASRU; 2007 Dec 9-13; Kyoto, Japan. p. 65-70.

Livescu K, Cetin O, Hasegawa-Johnson M, King S, Bartels C, Borges N, Kantor A, Lal P, Yung L, Bezman A, Bronwyn W. Articulatory feature-based methods for acoustic and audio-visual speech recognition: summary from the 2006 JHU summer workshop. In: IEEE, editor. Proc. ICASSP. 2007 Apr 15-20. Honolulu, Hawaii, USA. p. 621-4.

Lohscheller J, Eysholdt U, Toy H, Dollinger M. Phonovibrography: mapping high-speed movies of vocal folds vibrations into 2-D diagrams for visualizing and analysing the underlying laryngeal diseases. IEEE Trans. on Biomedical Eng. 2008;27(3):300-9.

Luengo I, Navas E, Hernaez I. Feature analysis and evaluation for automatic emotion identification in speech. IEEE Transactions on Multimedia. 2010;12(6):490-501.

Ma L, Milner B, Smith D. Acoustic environment classification. ACM Trans. Speech Lang. Process. 2006;3(2):1-22.

Maganti HK, Gatica-Perez D, McCowan I. Speech enhancement and recognition in meetings with an audio-visual sensor array. IEEE Transactions on Audio, Speech and Language Processing. 2007;15(8):2257-69.

Mariño Acebal J. Avivavoz: tecnologías para la traducción de voz. In: IV Jornadas en Tecnología del Habla. 2006 Nov 12-16; Zaragoza. Spain. p. 285-90.

Mariño JB, Banchs RE, Crego JM, de Gispert A, Lambert P, Fonollosa JAR, Costa-Jussà MR. N-gram-based machine translation. Computational Linguistics. 2006;32(4):527-49.

Meignier S, Moraru D, Fredouillea C, Bonastre J-F, Besacier L. Step-by-step and integrated approaches in broadcast news speaker diarization. Computer Speech and Language. 2006;20:303-30.

Miguel A, Lleida E, Rose R, Buera L, Saz O, Ortega A. Capturing local variability for speaker normalization in speech recognition. IEEE Transactions on Audio Speech and Language Processing. 2008;16(3):578.

Moore D, McCowan I. Microphone array speech recognition: experiments on overlapping speech in meetings. In: IEEE, editor. Proc. ICASSP; 2003 Apr 6-10; Hong Kong, China. p.V-497-V-500.

Moreno A. Information search engine for multilingual audiovisual contents: BUCEADOR. FALA 2010. 2010 Nov 10-12; Vigo, Spain. p. 259-62.

Moreno A, Poch D, Bonafonte A, Lleida E, Llisterri J, Marino JB, Nadeu C. Albayzin speech data base: design of the phonetic corpus. In: ISCA, editor. Proceedings of EUROSPEECH'93. 1993 Sep 21-23; Berlin, Germany. p.175-8.

Munkong R, Juang BH. Auditory Perception and Cognition. IEEE Signal Proc. Magazine. 2008;56(9):98-117.

Nadas A. On Turing's formula for word probabilities. IEEE Trans. Acoustics, Speech, and Signal Proc. 1984;33:1414-6.

Nadeu C, Macho C, Hernando J. Frequency & time filtering of filter-bank energies for robust HMM speech recognition. Speech Communication (Special Issue on Noise Robust ASR). 2001;34:93-114.

Nadeu C, Pachés-Leal P, Juang BH. Filtering the time sequences of spectral parameters for speech recognition. Speech Communication. 1997;21:315-32.

Navarro-Mesa J-L, Quintana-Morales P, Pérez-Castellano I, Espinosa-Yáñez J. Oral corpus of the project HACRO (help tool for the confidence of oral utterances) [technical report]. Las Palmas de Gran Canaria, Spain: University of Las Palmas de Gran Canaria; 2005.

Navas E, Hernáez I, Luengo I. An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS. IEEE Trans. Audio, Speech and Lang. Proc. 1997;14(4):1117-27.

Navas E, Hernaez I, Luengo I, Sanchez J, Saratxaga I. Analysis of the suitability of common corpora for emotional speech modelling in standard basque. Lecture Notes in Artificial Intelligence. 2005;3658:265-72.

Navas E, Hernáez I, Sánchez J. Basque intonation modelling for text to speech conversion. In: ISCA, editor. Proc. of 7th International Conference on Spoken Language Processing (ICSLP). 2002 Sep 16-20. Denver, USA. p. 2409-12.

Ntalampiras S, Potamitis I, Fakotakis N. On acoustic surveillance of hazardous situations. In: IEEE, editor. Proc. ICASSP. 2009 Apr 19-24; Taipei, China. p. 165-8.

Nguyen T, Sun H, Zhao S, Khine SZK, Tran HD, Ma TLN, Ma B, Chang ES, Li H. The IIR-NTU Speaker Diarization Systems for RT 2009 [accessed Jun 2011]. Available from: http://www.itl.nist.gov/iad/mig/tests/rt/2009/workshop/IIR-NTU-presentation.pdf

Och FJ. Minimum Error Rate Training in Statistical Machine Translation. In: 41st Annual Meeting of the Association for Computational Linguistics (ACL); 2003July 7-12; Sapporo, Japan. p.160-7.

Oester A-M, House D, Protopapas A, Hatzis A. Presentation of a new EU project for speech therapy: OLP (Ortho-Logo-Paedia). In: Proceedings of the XV Swedish Phonetics Conference (Fonetik 2002). 2002 May 29-31. Stockholm, Sweden. p. 45-8.

Ortega-García J, Fierrez J, Alonso-Fernandez F, Galbally J, Freire MR, Gonzalez-Rodriguez J, Garcia-Mateo C, Alba-Castro J-L, Gonzalez-Agulla E, Otero-Muras E, Garcia-Salicetti S, Allano L, Ly-Van B, Dorizzi B, Kittler J, Bourlai T, Poh N, Deravi F, Ng MWR, Fairhurst M, Hennebert J, Humm A, Tistarelli M, Brodo L, Richiardi J, Drygajlo A, Ganster H, Sukno FM, Pavani S-K, Frangi A, Akarun L, Savran A. The Multi-Scenario Multi-Environment BioSecure Multimodal Database (BMDB). IEEE Trans. on Pattern Analysis and Machine Intelligence. 2010;32(6):1097-111.

Osma-Ruiz V, Godino-Llorente JI, Sáenz-Lechón N, Fraile-Muñoz R. Segmentation of the glottal space from laryngeal images using the watershed transform. Computerized Medical Imaging and Graphics. 2008a;32:193-201.

Pardo JM, Anguera X, Wooters C. Speaker diarization for multiple distant-microphone meetings using several sources of information. IEEE Transactions on Computers. 2007;56(9):1212-24.

Peinado AM, Sánchez V, Pérez-Córdoba J, Rubio A. Efficient MMSE-based channel error mitigation techniques. Application to distributed speech recognition over wireless channels. IEEE Tr. Wireless Comm. 2005;4(1):14-9.

Peinado AM, Segura JC. Speech recognition over digital channels: robustness and standards. New York, USA: John Wiley & Sons Ltd.; 2006.

Perez-Freire L, Garcia-Mateo C. A multimedia approach for audio segmentation in TV broadcast news. In: IEEE, editor. Proc. of IEEE Internacional Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol 1. 2004 May 17-21; Montreal, Canada. p. 369-72.

PHAUST. Feedback analysis for user adaptive statistical translation. [Accessed Jun 2011] 2010. Available from: http://divf.eng.cam.ac.uk/faust

Pieraccini R, Levin E. A learning approach to natural language understanding. NATO-ASI, New Advances & Trends in Speech Recognition and Coding, Springer-Verlag, Bubion, Spain. 1993.

Pizarro D, Mazo M, Santiso E, Marron M, Fernandez I. Localization and geometric reconstruction of mobile robots using a camera ring. IEEE Transactions on Instrumentation and Measurement. 2009;58(8):2396-409.

Portelo J, Bugalho M, Trancoso I, Neto J, Abad A, Serralheiro A. Non-speech audio event detection. In: IEEE, editor. ICASSP 2009; Int. Conf. on Acoustics, Speech, and Signal Processing; 2009 Apr 19-24; Taiwan. China. p. 1973-1976.

Rabiner, LR, Wilpon, JG and Soong, FK. High performance connected digit recognition, using hidden Markov models. Proceedings of International Conference on Acoustic Speech, and Signal Processing, ICASSP-1988, 1988 Apr . 11-14, New York, USA. pp.119-122.

Rauschecker JP, Scott SK. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature Neuroscience. 2009;12(6):718-24.

Raux A, Black A. A unit selection approach to F0 modeling and its application to emphasis. In: Proc. ASRU. 2003 30 Nov.-3 Dec. St Thomas, US Virgin Is. p.700-5.

Rodríguez-Fuentes LJ, Peñagarikano M, Bordel G, Varona A, Díez M. KALAKA: A TV broadcast speech database for the evaluation of language recognition systems. In: Proc. LREC. 2010 May 17-23; Valletta. Malta. p. 1895-8.

Rosenfeld R. Adaptive statistical language modeling: a maximum entropy approach [Ph.D. thesis]. Pittsburgh, USA: Carnegie Mellon University; 1994.

Sáenz-Lechón N, Godino-Llorente JI, Osma-Ruiz V, Gómez-Vilda P. Methodological issues in the development of automatic systems for voice pathology detection. Biomedical Signal Processing and Control. 2006;1(2):120-8.

Sanchis E, Buscaldi D, Grau S, Hurtado L, Griol D. Spoken QA based on a passage retrieval engine. In: Proc. of IEEE/ACL 2006 Workshop on Spoken Language Technology (SLT). 2006 Dec 10-13; Aruba. p. 62-5.

Sanchos E, Segarra E; Torres F. User simulation in a stochastic dialog system. Computer Speech and Language. 2008;22:230-55.

San-Segundo R, Barra R, Córdoba R, D'Haro LF, Fernández F, Ferreiros J, Lucas JM, Macías-Guarasa J, Montero JM, Pardo JM. Speech to sign language translation system for Spanish. Speech Communication. 2008;50:1009-20.

San Segundo R, Montero JM, Guarasa JM, Ferreiros J, Pardo JM. Knowledge-combining methodology for dialogue design in spoken language systems. International Journal of Speech Technology. 2005;8(1):45-66.

Saz O. On line personalization and adaptation to disorders and variations of speech on automatic speech recognition systems [PhD thesis]. Zaragoza: Universidad de Zaragoza; 2009. Available from: http://dihana.cps.unizar.es/~oscar/data/Tesis_Oscar_Saz.pdf.

Saz O, Lleida E, Miguel A. Combination of acoustic and lexical speaker adaptation for disordered speech recognition. In: ISCA, editor. Proc. of the 11th European Conference on Speech Communication and Technology (Eurospeech-Interspeech); 2009 Sep 6-10; Brighton, United Kingdom. p. 544-7.

Saz O, Rodríguez W-R, Lleida E, Vaquero C. A novel corpus of children's impaired speech. In: Proceedings of the 2008 Workshop on Children, Computer and Interaction; 2008 Oct 23; Chania, Crete, Greece. p. 1-6.

Saz O, Yin S-C, Lleida E, Rose R, Rodríguez W-R, Vaquero C. Tools and technologies for computer-aided speech and language therapy. Speech Communication. 2009c;51(10):948-67.

Scherer KR. Vocal communication of emotion: a review of research paradigms. Speech Communication. 2003;40:227-56.

Seneff S. 1992 TINA: A natural language system for spoken language applications. Computational Linguistics. 1992;18(1):61-86.

Shriberg, E., Stolcke, A. & Baron, D. Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation. In: ISCA, editor. Proc. Conference EUROSPEECH. 2001 Sep 3-7. Aalborg, Denmark. p. 1359-62.

Yaman, S., Hakkani-Tür, D., Tur. G., Combining semantic and syntactic information sources for 5-W question answering. In: ISCA, editor. Proceedings of the Interspeech'09, Annual Conference of the International Speech Communication Association. 2009 Sep 6-10; Brighton, United Kingdom. p.2707-10.

van den Heuvel H, Boves L, Moreno A, Omologo M, Richard G, Sanders E. Annotation in the SpeechDat Projects. International Journal of Speech Technology. 2001;4(2):127-43.

Stylianou Y. Voice transformation: a survey. In: IEEE, editor. Proc. of the IEEE ICASSP. April 19-24; Tapei, Taiwan. 2009. p. 3585-8.

Taylor JG, Scherer K, Cowie R. Emotion and brain: understanding emotions and modelling their recognition. Neural Networks. 2005;18:313-6.

Temko A, Nadeu C. Acoustic event detection in a meeting-room environment. Pattern Recognition Letters. 2009;30(14):1281-8.

Temko A, Nadeu C, Macho D, Malkin R, Zieger C, Omologo M. Acoustic event detection and classification. In: Waibel A, Stiefelhagen R, editors. Computers in the human interaction loop. London: Springer; 2009. p. 61-73.

Torre-Toledano D, Lopez-Moreno I, Mateos I, Abejón A, Ramos D, Gonzalez-Rodriguez J. Automatic language recognition on spontaneous speech: the ATVS-UAM system. JAES Journal on Audio Engineering Society. 2009;10(57):788-806.

Toselli AT, Romero V, Pastor-i-Gadea M, Vidal E. Multimodal interactive transcription of text images. Pattern Recognition. 2010;43(5):1814-25.

Tranter S, Reynolds DA. An overview of automatic speaker diarization. IEEE Trans. On Audio, Speech and Language Processing. 2006;14(5):1557-65.

Turmo J, Comas PR, Rosset S, Galibert O, Moreau N, Mostefa D, Rosso P, Buscaldi D. Overview of QAST 2009. In: 10th Int. Cross-Language Evaluation Forum CLEF-2009 working notes; 2009 Sep 30-Oct 2; Corfu, Greece. p. 253-256.

Vaquero C, Ortega A, Lleida E. Intra-session variability compensation and hypothesis generation and selection strategy for speaker segmentation. In: IEEE, editor. International Conference on Acoustics, Speech and Signal Processing ICASSP; 2011 May 22-27; Prague, CZech Republic. p. 4532-5.

Vicente-Peña J, Gallardo-Antolín A, Peláez D, de María FD. Band-pass filtering of the time sequences of spectral parameters for robust wireless speech recognition. Speech Communication. 2006;48(10):1379-98.

Vicsi K, Roach P, Oester A, Kacic Z, Barczikay P, Sinka I. SPECO: A multimedia multilingual teaching and training system for speech handicapped children. In: ISCA, editor. Proc. of the 6th European Conference on Speech Communication and Technology (Eurospeech-Interspeech). 1999 Sep 5-9; Budapest, Hungary. p. 859-62.

Vijayasenan D, Valente F, Bourlard H. An information theoretic approach to speaker diarization of meeting data. IEEE Transactions on Audio, Speech, and Language Processing. 2009;17(7):1382-93.

Waibel A, Fügen C. Spoken language translation. IEEE Signal Processing Magazine. 2008;25(3):70-9.

Wang Y. A robust parser for spoken language understanding. In: ISCA, editor. Proc. of EUROSPEECH. 1999 Sep 5-9; Budapest, Hungary. p. 2055-8.

Wang Y, Acero A. Grammar learning for spoken language understanding. In: Proceedings of IEEE ASRU Workshop; 2001 Dec 9-13; Madonna di Campiglio, Italy. p. 1229-44.

Wang Y, Acero A, Chelba C, Frey B, Wong L. Combination of statistical and rule-based approaches for spoken language understanding. In: ISCA, editor. Proc. of ICSLP; 2002 Sep 16-20; Denver, Colorado, USA. p. 609-12.

Ward W, Issar S. Recent improvements in the CMU spoken language understanding system. In: Proceedings of ARPA Workshop on HLT. 1994 Mar 8-11; Plainsboro, New Jersey, USA. p. 213-6.

Wooters C, Huijbregts M. The ICSI RT07s speaker diarization System. In: Proceedings of the Rich Transcription 2007 Meeting Recognition; 2007 May 1-4; Baltimore, Maryland. p. 509-19.

Yaman S, Hakkani-Tür D, Tur G, Grishman R, Harper M, McKeown KR, Meyers A, Sharma K. Classification-based strategies for combining multiple 5-W question answering systems. In: ISCA, editor. Proceedings of the Interspeech'09, Annual Conference of the International Speech Communication Association. 2009 Sep 6-10; Brighton, UK. p. 2703-6.

Yamagishi J, Nose R, Zen H, Ling A, Toda T, Tokuda K, King S, Renals S. A robust speaker-adaptive HMM-based text-to-speech synthesis. IEEE Trans. Audio, Speech and Lang. 2009;17(6):1208-30.

Yan Y, Chen X, Bless D. Automatic tracing of vocal-fold motion from high-speed digital images. IEEE Trans. on Biomedical Eng. 2006;53(7):1394-400.

Zang Y, Bieginga, E., Tsuia H., and Jiang J., Efficient and effective extraction of vocal fold vibratory patterns from high-speed digital imaging. Journal of Voice. 2010;24(1):21-9.

Zen H, Tokuda K, Black A. Statistical parametric speech synthesis. Speech Communication. 2009;51(11):1039-54.