applied sciences

MDPI

Editorial

Editorial for Special Issue "IberSPEECH2018: Speech and Language Technologies for Iberian Languages"

Francesc Alías [1,*,†], Antonio Bonafonte [2,†] and António Teixeira [3,4,†]

1 GTM—Grup de Recerca en Tecnologies Mèdia, La Salle—Universitat Ramon Llull, Quatre Camins 30, 08022 Barcelona, Spain
2 Amazon, Cambridge CB1 2GA, UK; bonafont@amazon.com
3 Department of Electronics Telecommunications and Informatics, University of Aveiro, 3810-193 Aveiro, Portugal; ajst@ua.pt
4 Biomedical Informatics and Technologies Group, IEETA, 3810-193 Aveiro, Portugal
* Correspondence: francesc.alias@salle.url.edu; Tel.: +34-932-902-440
† These authors contributed equally to this work.

Received: 19 December 2019; Accepted: 30 December 2019; Published: 4 January 2020

Abstract: The main goal of this Special Issue is to present the latest advances in research and novel applications of speech and language technologies based on the works presented at the IberSPEECH edition held in Barcelona in 2018, paying special attention to those focused on Iberian languages. IberSPEECH is the international conference of the Special Interest Group on Iberian Languages (SIG-IL) of the International Speech Communication Association (ISCA) and of the Spanish Thematic Network on Speech Technologies (Red Temática en Tecnologías del Habla, or RTTH for short). Several researchers were invited to extend their contributions presented at IberSPEECH2018 due to their interest and quality. As a result, this Special Issue is composed of 13 papers that cover different topics of investigation related to perception, speech analysis and enhancement, speaker verification and identification, speech production and synthesis, natural language processing, together with several applications and evaluation challenges.

Keywords: Iberian languages; speech production; speech synthesis; speech recognition; speaker identification; speech enhancement; summarization; semantic representations; natural language processing; neural networks; deep learning; evaluation challenge; audiovisual database

1. Introduction

Several languages originated in theR Iberian peninsula and are known as Iberian Languages. The set includes Spanish, Portuguese, Catalan, Galician, Basque and other less known languages (e.g., Mirandese or Aragonese). Together, they are now the first language for more than 750 million speakers [1] and two of them, Spanish and Portuguese, are in the top 10 of the most spoken languages in the world [1].

Research targeting their processing has a long tradition, with several well-known research groups and researchers. A process initiated in 2005 resulted in the creation of a Special Interest Group of the International Speech Communication Association (ISCA), named SIG-IL, aiming to promote research activities and research interests in Iberian Languages, in both spoken and written forms. As other ISCA special interest groups on specific languages (e.g., Chinese, Russian, Italian, French), SIG-IL organizes an international conference, the IberSPEECH that aims to promote the presentation of recent advances in the field and increase interaction and discussion among the members of this research community.

The IberSPEECH series of conferences has become one of the most relevant scientific events for the community working in the field of speech and language processing of Iberian languages. This is

demonstrated by the increased interest and success of previous editions, starting in Vigo (2010), with FALA, and continuing, with the current name, in Madrid (2012), Las Palmas de Gran Canaria (2014), and Lisbon (2016), with the last edition in Barcelona (2018).

The main goal of this Special Issue is to present the latest advances in research and novel applications of speech and language processing, paying special attention to those focused on Iberian languages, based in research presented at IberSPEECH2018, edition held in Barcelona, Spain. For this first ever Special Issue resulting from an IberSPEECH conference, a very rich and diverse set of high-quality papers were selected. They are briefly presented in the next section, organized by research topics.

2. Contributions

2.1. Speech Analysis and Enhancement

The selection includes two papers related to speech analysis and enhancement, a traditional key topic in the area of speech processing, which are focused on the automatic discrimination of singing from speech, and the improvement of speech recognition systems running on two-microphone smartphones, respectively.

The article by Sarasola et al. [2] introduces an automatic classification system capable of discriminating speech from singing voice, using two parameters derived from their fundamental frequency. This research derives from the need to separate speech segments from a capella singing segments recorded in Bertsolarism (Bertsolaritza in Basque), the traditional art of live improvising sung poems in the Basque Country. The proposal is developed on the Bertsolaritza database, and it is subsequently tested on this dataset and on the English Sung and Spoken Lyrics Corpus from the National University of Singapore. The proposal combines a Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) based voice activity detection, a pitch-based Note Detection Algorithm (as a by-product of the discrimination approach) and a binary GMM-based classifier that differentiates speech from singing. The system obtains good results when discriminating both voice types, obtaining a better performance than the baseline systems based on pitch parameters, and similar results as the spectrum analysis system. Finally, the proposal shows good generalization capabilities, besides entailing a low computation time.

The article by Martín-Doñas et al. [3] deals with speech enhancement in dual-microphone smartphones using beamforming along with post-filtering techniques. In this context, it is worth mentioning that the performance of beamforming algorithms is diminished on dual-microphone smartphones mainly due to the short number of microphones, the close separation between them and their placement on the device. In order to improve the estimation of the acoustic channel and the main characteristics of speech and noise when using the Minimum Variance Distortionless Response (MVDR) beamformer, the authors propose combining the estimation of the relative transfer function (RTF) between microphones using an extended Kalman filter (eKF) together with a speech presence probability (SPP) estimator intended to track the background noise variability. Moreover, the proposal also includes postfiltering that make use of the previously estimated parameters to design a gain function which is applied to the beamformer output signal for further noise reduction. The experiments evaluate the proposal in different reverberant and noisy environments, considering both close-talk and far-talk environments and simulating dual-channel noisy speech recordings from a dual-microphone smartphone positions using the English TIMIT database as input. The results show that the proposal improves the estimation of RTFs and SPPs with respect to state-of-the-art algorithms, since it yields lower speech distortion and better speech quality and intelligibility. Moreover, the proposed postfilters also are capable of improving the noise reduction of the MVDR-based output compared to those algorithms specifically intended for dual-microphone smartphones without degrading the speech intelligibility.

2.2. Speaker Verification and Identification

Speech classification and identification is a research area that has consistently been highly represented in IberSPEECH conferences, including the 2018 edition. This special issue includes two papers from ViVoLab [4,5]. In these works, the authors investigate how to use phonetic information for the tasks of short sentence verification and text-independent speaker verification using Neural Networks. Moreover, the paper from Khan et al. [6] explores the use of Boltzmann Machines to speaker clustering and tracking. Furthermore, the research conducted by Rituerto-González et al. [7] apply speaker identification techniques in a real application under adverse conditions (see Section 2.6).

Viñals et al. [4] analyze the problem of short sentences in Speaker Characterization and, in particular, in Speaker Identification. It is well known that the performance of these systems degrades with short sentences. Most of state-of-the-art systems, including i-vectors and Deep Neural Networks (DNN) vectors, apply a scoring function based on test and reference vectors. The authors observe that these vectors can be expressed as the weighted average of the vectors describing each phoneme. A meaningful comparison between test and reference vectors requires a good estimation not only of the phoneme characterization, but also of the relative frequency of each phoneme. In fact, they show that in short sentences the mismatch in phoneme frequencies between test and reference is the main cause of error. They show that short utterances (3–60 s) can perform almost with the same accuracy than long sentences if there is a match between phonetic distribution.

Mingote et al. [5] propose an architecture to include phonetic information in text-dependent verification systems using Neural Networks. DNN has improved significantly many speaker verification tasks. However, these models do not offer the same boost in performance in the case of text-dependent verification systems, when the text is part of the identity information. The problem with traditional DNN is that the pooling mechanism used to transform the acoustic sequence in a fixed-dimension vector does not preserve the phonetic order, therefore disregarding this important source of information. This paper combines DNN with an external alignment which can be computed using HMM-based forced alignment. The model first applies 1D convolutional networks to get a more powerful representation of the frames. Then, the frames associated with the same HMM-state are averaged to produce a fixed-length vector. Finally, a projection layer reduces the vector dimension and the cosine metric is applied. This architecture achieves much better results than the traditional average, resulting in competitive results.

Khan et al. [6] apply features derived using the Restricted Bolzmann machine (RBM) to the task of speaker clustering and tracking. Previously, the authors had successfully shown that RBM can offer competitive results in speaker recognition. Similarly, in this paper, the speaker is represented using a RBM vector, which is derived applying PCA whitening to the visible to hidden weight matrices of the RBM. For speaker clustering, an Agglomerative Hierarchical Clustering algorithm is applied and the results in terms of Equal Impurity show a relative 11% improvement with respect to i-vectors. Similarly, using RBM vectors in the identification step of a speaker tracking algorithm, the results are also better than i-vectors both using cosine distance or PLDA as scoring function.

2.3. Speech Production and Synthesis

Speech synthesis research for the several Iberian languages has a long tradition, with systems being developed based both in state-of-the-art technologies as well on methods closer to human speech production. For this special issue, very good representatives of both lines of research were selected: one using recent neural architectures for Linguistic-Acoustic Mapping [8] and the other exploring finite element synthesis of vowels [9].

Current state-of-the-art Text-to-Speech systems use statistical models such as deep neural networks to convert text to speech both in end-to-end and two-stage variants. Recurrent neural networks are a specific neural topology that proved capable of producing high quality speech, but the high computation demands, caused by the recurrent connections, make them unsuitable for low resource environments such as embedded devices or to run locally in small and mobile

devices (e.g., smartphones). The first paper of this section [8], by Pascual, Serrà, and Bonafonte, focusing on the second part of a two-stage Text-to-Speech system, investigates both speech quality and computational efficiency of two competitive processing architectures: a linguistic-acoustic decoder based on quasi-recurrent neural networks (QLADs) and a self-attention linguistic acoustic-decoder (SALAD), based on the Transformer network. From the results of objective and subjective evaluation, performed with the Spanish speech dataset, QLAD proved to be a high quality and computational efficient replacement for recursive solutions.

The second paper of this section is also related to acoustic modeling for speech synthesis, but this time the model is directly inspired by human speech production structures and physical phenomena and only for vowels. The paper, by Freixes and coworkers [9], adopts realistic 3D acoustic models to include the higher order models of propagation that typically appear above 5 kHz to study the changes in high frequency. Their work is motivated by the limitation of 1D approaches to planar propagation and the little attention of previous research to the high frequency range. The authors experimented with two vocal tract representations (realistic and simplified) derived from MRI data with glottal source configurations covering the whole phonation range. Analysis of the long-term average spectrum of the obtained synthetic speech showed diminished high frequency energy for all phonation types and all the vowels synthesized. This reduction, between 3.6 dB and 7.2 dB, strongly dependent on phonation type and fundamental frequency, may be perceptually relevant.

2.4. Natural Language Processing

Despite speech related topics having a preponderance in the community of researchers contributing to IberSPEECH, the conference, aligned with the evolution of major conferences such as IEEE ICASSP and ISCA InterSpeech, includes as part of its main topics Natural Language Processing, as well as several topics requiring combination of speech and language (e.g., dialogue systems). For this special issue, two articles were select.

The first one, by González et al. [10], addresses the increasing problem of dealing with the vast amount of information available by adopting automatic summarization. The paper focus is the summarization of TV programs transcriptions, using a system based on Hierarchical Attention Networks trained in a Siamese way. Two corpora were built: one of news article-summary pairs, for training; the other, for evaluation, with pairs of transcriptions of TV programs and associated summary. Results of the system with news corpus are similar to the ones obtained for English with a similar neural architecture. When tested with TV transcriptions, the proposed system obtained better evaluation metrics than a set of five representative extractive unsupervised summarization systems.

The second paper of the section, "Towards a Universal Semantic Dictionary" [11], by Castro-Bleda and co-authors, is part of the recent direction in the field of computational semantics of investigating universal meaning representations. As meaning representation, the state-of-the-art solution of word embeddings is adopted. The novel method proposed, based in the usual approach in the field of using transformation matrices to map an embedding from one language to another, is capable of learning linear mappings among word embeddings of several languages. Evaluation results with pairs of languages, using established standards and new corpora created by the authors, are comparable with more sophisticated systems. The system variant trained on three different languages at the same time, and involving multilingual representations, showed a lower performance, but promising and pointing to the viability of learning multilingual representations.

2.5. Perception

Speech and language research is not only technologies and their applications. Improved knowledge will potentiate technologies and guide future research efforts. In this line, the article by Raman et al. [12] contributes to a better understanding of the communication challenges faced by oesophageal speakers when interacting with other humans and machines, investigating Intelligibility and Listening Effort in comparison to healthy laryngeal speech. They observed that,

despite intelligibility of oesophageal speech can be close to healthy laryngeal speech, it implies a greater listening effort, somewhat reduced for those familiar with it. Results justify the authors' ongoing research in oesophageal speech restoration.

2.6. Applications of Speech Technologies

Although several papers of the Special Issue are focused somehow on specific technical research issues, this section summarizes two investigations oriented to specific applications—the first one, designed to help people with Down syndrome to improve speaking pronunciation, and the second one focuses on combating the scourge of gender-based violence using speech technologies.

The article by Corrales-Astorgano et al. [13] aims at developing an automatic classifier to predict the prosodic quality of utterances produced by individuals with Down syndrome. To that effect, it considers a classic set of speech features and a supervised classification approach, but paying special attention to the inter-individual heterogeneity of the speakers according to the results obtained in the Spanish version of the Profiling Elements of Prosody in Speech-Communication (PEPS-C) test. In order to collect enough representative speech data, the authors have developed an educational video game to train prosody implemented as a graphic adventure. The appropriateness of the corpus were evaluated by a therapist and an expert on speech technologies, and this information was used to train the classifiers. Although the best supervised classification approach yields around 80% of accuracy, the authors highlight the high dependence of the results on the developmental level of the speakers, an issue that should be considered to build automatic evaluation systems accounting for this typical variability in the cognitive and linguistic skills of individuals with Down syndrome.

The article by Rituerto-González et al. [7] deals with the development of a speaker identification robust to emotional bias. The system is designed to run on a personalized wearable pendant, which is being developed within the Bindi project as a smart solution to improve women's safety. Specifically, the work focuses on addressing the problem of data scarcity to train speaker recognition systems to effectively identify the speaker identity under emotional or stress conditions. Due to inherent complexity to collect this kind of data in real-life conditions, the authors analyze the viability of applying synthetic data augmentation techniques to address the problem at hand. The experiments have been conducted using the VOCE Corpus Database in Portuguese, which contains similar data to what will be obtained by the Bindi system. The results show that the best performances are obtained when naturally stressed samples are included in the training dataset together with neutral speech (96% of accuracy). However, the authors also highlight that their substitution and augmentation with synthetically generated stressed speech (from neutral or naturally stressed speech) can improve the performance of the system when naturally stressed samples are not available in different degrees.

2.7. Evaluation

Since the first edition, IberSPEECH has included the *Albayzin* Evaluation Challenges in different speech technologies such as Spoken Term Detection, Language Identification, or Speech Synthesis. This special issue includes the paper from Lleida et al. [14], describing the IberSpeech-RTVE Challenge, organized by Radio Televisión Española, RTVE, the Spanish Public Broadcast Corporation and Vivolab, the speech research group at the University of Zaragoza. This Challenge consists of a set of technological evaluations in the areas of speech recognition and speaker diarization, including multimodal (audio and video) diarization.

First of all, the paper describes the RTVE 2018 database, a freely-available audiovisual database released for this evaluation by RTVE and the RTVE Chair at the University of Zaragoza. We believe that it will contribute significatively to the progress of Speech Technologies in Spanish, including the dialectal varieties spoken in Latin America. It contains more than 500 h of audio amb subtitles of broadcast data, spanning a broad range of genres. For this challenge, the database is a significant part of the database includes human-revised transcriptions and speaker labels.

The speech-to-text challenge has been evaluated using 39 h of audio from eight different TV shows, covering a range variety of acoustic conditions and speaker styles. For training, two scenarios were defined. In the close-set scenario, the systems could use approximately 70 h of transcribed speech and 460 h of untranscribed audio that includes subtitles. In the open-set scenario, participants were free to include any data to train their systems. Seven teams participated in the evaluation. The systems trained in the open-set scenario achieved in most of the systems significantly better results. The best Word Error Rate (WER) for open-set condition was 3% smaller than in close-set condition (16.45% vs. 19.57%). It was achieved by hybrid DNN-HMM systems.

For the speaker diarization challenge, the test data consisted of 22 h from different shows. For the close-set scenario, the same training data described above could be used. It includes 16 h of diarization annotations tracking the identity of the active speaker. The training speech database was extended with more than 100 h from other broadcast media. As in the other challenges, in the open-set scenario, any external data can be used to train the system. In this challenge, the results are similar for both open-set and close-set scenarios. The diarization error of the best system was 11.4% in the open-condition. In this scenario, the best system was much better than the other five participants. The analysis of the results show the huge impact on this metric of having a good estimation of the number of speakers present in the data.

The multimodal diarization challenge provided not only the speech, but also the video track along with enrollment audio, video, and pictures of the speakers to track. The training data included 2 h of annotated video, where the face of the speaker was identified, and 14 h with the identification of the active speaker. For the test, the system was provided 10 photos and 20 s of video of 39 characters. The smallest multimodal diaritzation error, defined as the mean between face and speaker diarization error, is approximately 23%. Without using the video information, the diarization error is 6% worse (29%). As in speech recognition, there are notable differences between shows, as some shows include frequent speaker overlap and exterior shots.

3. Discussion and Conclusions

This special issue has presented relevant recent advances of speech and language processing. Regarding the languages considered by these investigations, the majority of works have presented language-independent contributions. However, it is worth mentioning that several works have been considered Iberian Languages specifically or for evaluation purposes. Among them, Spanish has been considered by five papers, including one of them being a South American variant. Moreover, Catalan, Basque, and Portuguese have been used in one paper each. Regarding non-Iberian languages, five papers have considered English and one the Italian language.

During the last decade, Deep Learning has shifted the dominant techniques applied to many areas of machine learning, including speech and natural language processing. On one side, the increase of data and computation resources, along with advances in optimization, allows for building larger and more powerful models, which in many cases outperform recent state-of-the-art models. Furthermore, there has been a significant shift towards knowledge sharing, using not only traditional research papers but also through social media, blogs, and sharing of algorithms in the form of source code, data, and experiments. As a result, the number of publications in speech and language processing using deep learning techniques is more and more frequent in all the relevant journals and conferences. IberSPEECH has not been an exception, with more than 65% of the regular papers proposing Deep Learning based techniques to different problems. Deep Learning also plays a relevant role in this special issue, with 5 of 12 regular papers applying these techniques and most of the systems described in the Evaluation Challenge paper. However, the other seven papers show that there is still space for research in many relevant areas, as perception, signal processing, and application. Some of these research works suffer data sparsity which limits the direct application of the main trend deep-learning models. In other papers, the goal is understanding some of the fundamentals of perception and production.

Finally, we want to highlight the variety of research topics covered by this special number, which, together with the quality of the selected works, demonstrate the wide scope and quality of the research groups dedicated to Iberian languages, which will meet again to share the last advances in the forthcoming IberSPEECH2020 conference.

Author Contributions: Conceptualization, F.A., A.B., and A.T.; writing—original draft preparation, F.A., A.B., and A.T.; writing—review and editing, F.A., A.B., and A.T. All authors have read and agreed to the published version of the manuscript.

Funding: This special issue has been partially funded by the IberSPEECH2018 conference.

Acknowledgments: The authors would like to acknowledge the support of the Technical Program Committee of IberSPEECH2018 on the selection of the conference papers invited to this Special Issue and the Spanish Thematic Network on Speech Technologies (RTTH).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

eKF	extended Kalman Filter
DNN	Deep Neural Networks
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ICASSP	International Conference on Acoustics, Speech, and Signal Processing
IEEE	Institute of Electrical and Electronics Engineers
ISCA	International Speech Communication Association
PEPS-C	Profiling Elements of Prosody in Speech-Communication
MVDR	Minimum Variance Distortionless Response
MRI	Magnetic Resonance Imaging
PCA	Principal Component Analysis
PLDA	Probabilistic Linear Discriminant Analysis
QLAD	Quasi-recurrent neural networks Linguistic–Acoustic Decoder
RBM	Restricted Bolzmann machine
RTF	Relative Transfer Function
RTVE	Radio Televisión Española
SALAD	Self-Attention Linguistic Acoustic-Decoder
SPP	Speech Presence Probability
SIG-IL	Special Interest Group on Iberian Languages
RTTH	Spanish Thematic Network on Speech Technologies

References

1. Eberhard, D.M.; Simons, G.F.; Fennig, C.D. (Eds.) *Ethnologue: Languages of the World*, 22nd ed.; SIL International: Dallas, TX, USA, 2019.
2. Sarasola, X.; Navas, E.; Tavarez, D.; Serrano, L.; Saratxaga, I.; Hernaez, I. Application of Pitch Derived Parameters to Speech and Monophonic Singing Classification. *Appl. Sci*, **2019**, *9*, 3140. [CrossRef]
3. Martín-Doñas, J.M.; Peinado, A.M.; López-Espejo, I.; Gomez, A. Dual-Channel Speech Enhancement Based on Extended Kalman Filter Relative Transfer Function Estimation. *Appl. Sci.* **2019**, *9*, 2520. [CrossRef]
4. Viñals, I.; Ortega, A.; Miguel, A.; Lleida, E. An Analysis of the Short Utterance Problem for Speaker Characterization. *Appl. Sci.* **2019**, *9*, 3697. [CrossRef]
5. Mingote, V.; Miguel, A.; Ortega, A.; Lleida, E. Supervector Extraction for Encoding Speaker and Phrase Information with Neural Networks for Text-Dependent Speaker Verification. *Appl. Sci.* **2019**, *9*, 3295. [CrossRef]
6. Khan, U.; Safari, P.; Hernando, J. Restricted Boltzmann Machine Vectors for Speaker Clustering and Tracking Tasks in TV Broadcast Shows. *Appl. Sci.* **2019**, *9*, 2761. [CrossRef]

7. Rituerto-González, E.; Mínguez-Sánchez, A.; Gallardo-Antolín, A.; Peláez-Moreno, C. Data Augmentation for Speaker Identification under Stress Conditions to Combat Gender-Based Violence. *Appl. Sci.* **2019**, *9*, 2298. [CrossRef]
8. Pascual, S.; Serrà, J.; Bonafonte, A. Exploring Efficient Neural Architectures for Linguistic–Acoustic Mapping in Text-To-Speech. *Appl. Sci.* **2019**, *9*, 3391. [CrossRef]
9. Freixes, M.; Arnela, M.; Socoró, J.; Alías, F.; Guasch, O. Glottal Source Contribution to Higher Order Modes in the Finite Element Synthesis of Vowels. *Appl. Sci.* **2019**, *9*, 4535. [CrossRef]
10. González, J.A.; Hurtado, L.F.; Segarra, E.; García-Granada, F.; Sanchis, E. Summarization of Spanish Talk Shows with Siamese Hierarchical Attention Networks. *Appl. Sci.* **2019**, *9*, 3836. [CrossRef]
11. Castro-Bleda, M.J.; Iklódi, E.; Recski, G.; Borbély, G. Towards a Universal Semantic Dictionary. *Appl. Sci.* **2019**, *9*, 4060. [CrossRef]
12. Raman, S.; Serrano, L.; Winneke, A.; Navas, E.; Hernaez, I. Intelligibility and Listening Effort of Spanish Oesophageal Speech. *Appl. Sci.* **2019**, *9*, 3233. [CrossRef]
13. Corrales-Astorgano, M.; Martínez-Castilla, P.; Escudero-Mancebo, D.; Aguilar, L.; González-Ferreras, C.; Cardeñoso Payo, V. Automatic Assessment of Prosodic Quality in Down Syndrome: Analysis of the Impact of Speaker Heterogeneity. *Appl. Sci.* **2019**, *9*, 1440. [CrossRef]
14. Lleida, E.; Ortega, A.; Miguel, A.; Bazán, V.; Pèrez, C.; Gómez, M.; de Prada, A. Albayzin 2018 Evaluation: The IberSpeech-RTVE Challenge on Speech Technologies for Spanish Broadcast Media. *Appl. Sci.* **2019**, *9*, 5412. [CrossRef]