INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Compensating Gender Variability in Query-by-Example Search on Speech using Voice Conversion

Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo

AtlantTIC Research Center, Multimedia Technologies Group, University of Vigo, Spain
{plopez,ldocio,carmen}@gts.uvigo.es

Abstract

The huge amount of available spoken documents has raised the need for tools to perform automatic searches within large audio databases. These collections usually consist of documents with a great variability regarding speaker, language or recording channel, among others. Reducing this variability would boost the performance of query-by-example search on speech systems, especially in zero-resource systems that use acoustic features for audio representation. Hence, in this work, a technique to compensate the variability caused by speaker gender is proposed. Given a data collection composed of documents spoken by both male and female voices, every time a spoken query has to be searched, an alternative version of the query on its opposite gender is generated using voice conversion. After that, the female version of the query is used to search within documents spoken by females and vice versa. Experimental validation of the proposed strategy shows an improvement of search on speech performance caused by the reduction of gender variability.

Index Terms: query-by-example search on speech, dynamic time warping, voice conversion, variability compensation

1. Introduction

The amount of available spoken documents is steadily increasing, which leads to the need for tools to perform automatic searches within large audio databases. These data collections usually comprise documents of diverse sources with a great variability regarding content, speaker, language or recording conditions, among others. Searching within this type of databases using written keywords requires knowledge about the language being spoken, which makes these approaches unpractical for under-resourced languages and multilingual environments. Hence, the use of spoken queries has gained the attention of the research community, since these strategies allow to approach this task as a pattern matching problem.

In query-by-example search on speech (QbESOS), first a set of features is extracted from the queries and documents, and then a matching between the queries and documents is performed in order to find occurrences of the queries in the documents, which is usually carried out by means of dynamic time warping (DTW) algorithm [1] or any of its variants. The flexibility of these methods has attracted the interest of the research community, leading to the organization of international competitions [2, 3, 4, 5, 6, 7, 8]. The use of cross-lingual approaches such as phoneme posteriorgram representation [9] is common in QbESOS, as they enable the use of acoustic models available for one language when performing a search in spoken documents in a different language [10, 11]. In the last years, the interest in zero-resource QbESOS approaches has raised since they do not require any modelling, resources or knowledge about any language. In zero-resource QbESOS, the use of acoustic features such as Mel-frequency cepstral coefficients (MFCCs) or perceptual linear predictive coefficients (PLPs), which are typical in other speech processing tasks, is quite common, with or without further processing of the features [12, 13, 14]. The extraction of a large set of features and a posterior selection of those more relevant for the task has also been proposed [15], leading to an improvement of the results achieved using standard speech recognition features such as MFCCs.

It is straightforward to believe that QbESOS systems based on acoustic features are very sensitive to data variability. Given that such features are used for tasks such as speaker identification or gender classification, it seems obvious that they retain information about those characteristics which might act as nuisance when performing search on speech, since all the information regarding the speaker is irrelevant for this task. This suggests that a query would be easily found in a document spoken by that person or by another with a similar voice.

In this paper, a technique to compensate the variability in QbESOS is proposed, which aims at searching for queries within documents with similar voices. The approximation presented in this paper reduces the problem to searching for queries in documents spoken by people of the same gender. Since a query spoken by a female might appear in a document spoken by a male and *vice versa*, the use of voice conversion to modify the gender of the speaker is proposed in this work. Given that most voice conversion techniques require a parallel corpus between the source and target speaker to train the conversion function, the voice conversion technique presented in [16] was used, because it does not present that limitation.

The proposed technique for gender variability compensation was assessed in the evaluation framework defined for the Query by Example Search on Speech task at MediaEval 2014 (QUESST 2014) [6]. This experimental framework consists of a large collection of documents and queries in multiple languages and recording conditions, and the goal of the task is to perform spoken document retrieval: for each document-query pair, a score must be assigned such that the higher the score, the higher the chance that the query was pronounced within the document. The experimental validation showed that the proposed technique for gender variability compensation improved the results with respect to the reference system. In addition, a comparison was also established with a system using log-likelihood scores of a gender classifier as side information [17], but this approach was also outperformed by the technique proposed in this paper.

The rest of this paper is organized as follows: Section 2 presents the proposed gender variability compensation approach for QbESOS; the experimental framework used to validate this technique is described in Section 3; Section 4 presents the experimental results; and, finally, some conclusions and future work are summarised in Section 5.

http://dx.doi.org/10.21437/Interspeech.2017-1183

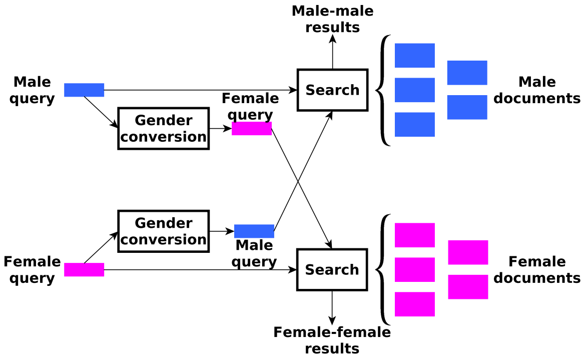

Figure 1: *Block diagram of the proposed approach for gender variability compensation.*

2. Gender variability compensation in QbESOS

Figure 1 presents an overview of the proposed technique for gender variability compensation in a QbESOS scenario. First, the database of documents must be classified by gender. After that, when a spoken query is fed to the system, its gender must be detected and, in case it is a male query, it is transformed into a female one using voice conversion and vice versa. After that, the male and female queries are searched within the male and female documents, respectively. Hence, the proposed system comprises three different stages: gender classification, gender conversion, and search. The rest of this Section describes the implementation for these blocks used in this work.

2.1. Gender classification

In this paper, a gender classifier based on Gaussian mixture model (GMM) log-likelihood ratio was used. Given male and female GMMs, the likelihood of a test utterance given each GMM is computed and their log-likelihood ratio is calculated, assigning the most likely gender to that utterance [18].

It must be noted that using such approach for gender detection might lead to classification errors, but gender detectors usually have a classification performance close to 100% and, in addition, given the nature of the approach presented here, the apparent gender of the voice is more relevant than is actual gender. In other words, if the speaker of a query is a male with a high-pitched voice, performing a female-to-male conversion would be more convenient than doing a male-to-female conversion, since the latter would result in a rather high-pitched voice.

2.2. Gender conversion

The gender of a speaker is mainly determined by the fundamental frequency (F0) and formant frequencies of their voice [19]. Modifying these parameters in the proper way results in a change of the perceived speaker gender, and this can be accomplished by means of voice conversion. This technique consists in modifying the voice characteristics of a source speaker in order to make it sound like a target speaker. Typical voice conversion techniques require a parallel corpus to train the transformation functions, which is not available in this scenario. Hence, the technique proposed in [16] was used in this work. This approach is based on frequency warping (FW) combined with amplitude scaling (AS), which consists in applying a linear transform in the cepstral domain [20]:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{1}$$

where $\mathbf{x}$ is a Mel-cepstral vector, $\mathbf{A}$ denotes a FW matrix, $\mathbf{b}$ represents an AS vector, and $\mathbf{y}$ is the transformed version of $\mathbf{x}$.

As mentioned before, traditional FW+AS strategies perform a training stage in order to obtain the transformation parameters to turn a given source speaker into its corresponding target speaker. In the method proposed in [16], the FW curve is simplified and defined piecewise by means of three linear functions. Hence, there are some parameters to be defined, which are: the frequencies where the discontinuities are produced $\mathrm{f_a}$ and $\mathrm{f_b}$; the angle α between the 45-degree line and the first linear function; and the angle β between the 45-degree line and the second linear function, which is defined as $\beta = \mathrm{k}\alpha$ $(0 < \mathrm{k} < 1)$. Values of α greater than 0 move the formants to higher frequencies, so they result in a transformation function suitable for performing male-to-female conversion. On the contrary, negative values of α yield suitable female-to-male conversion functions.

The AS vector $\mathbf{b}$ is defined by randomly giving values to a set of weighted Hanning-like bands equally spaced in the Mel-frequency scale [21] as fully described in [16]. Finally, FW+AS is complemented with a scaling of the fundamental frequency proportional to the value of α [16].

2.3. Search

The strategy for QbESOS described in [15] was employed in this paper. It comprises three stages: feature extraction, search, and score normalisation. Details on the system are given below.

First, a large set of features, which is summarised in Table 1, was extracted from the queries and the documents using the OpenSMILE toolkit [22]. After that, the feature selection procedure proposed in [23] was performed in order to keep only those features that are relevant for the task. This technique relies on the contribution of each feature to the best alignment path in terms of correlation, and has shown to improve performance in zero-resource QbESOS experiments [15].

After feature extraction, given a query $\mathrm{Q} = \{\mathbf{q_1}, \ldots, \mathbf{q_n}\}$ and a document $\mathrm{D} = \{\mathbf{d_1}, \ldots, \mathbf{d_m}\}$ of n and m frames respectively, with vectors $\mathbf{q}_i$ and $\mathbf{d}_j$ of F features and $\mathrm{n} \ll \mathrm{m}$, DTW finds the best alignment path between these two sequences. Among the available variants of DTW, subsequence DTW [24] (S-DTW) was used in this system, since it allows the alignment of a short sequence with a part of a longer sequence. First, a cost matrix $\mathbf{M} \in \Re^{\mathrm{n}\times\mathrm{m}}$ is defined such that its rows correspond to the frames of the query and its columns correspond to the frames of the document:

$$\mathrm{M}_{i,j} = \begin{cases} \mathrm{c}(\mathbf{q}_i, \mathbf{d}_j) & \text{if} \quad i = 0 \\ \mathrm{c}(\mathbf{q}_i, \mathbf{d}_j) + \mathrm{M}_{i-1,0} & \text{if} \quad i > 0 \\ & \qquad\quad j = 0 \\ \mathrm{c}(\mathbf{q}_i, \mathbf{d}_j) + \mathrm{M}^*(i,j) & \text{else} \end{cases} \tag{2}$$

where $\mathrm{c}(\mathbf{q}_i, \mathbf{d}_j)$ is a function that defines the cost between query vector $\mathbf{q}_i$ and document vector $\mathbf{d}_j$, and

$$\mathrm{M}^*(i,j) = min\left(\mathrm{M}_{i-1,j}, \mathrm{M}_{i-1,j-1}, \mathrm{M}_{i,j-1}\right) \tag{3}$$

Pearson's correlation coefficient r is used as cost function in this system [25] since it empirically showed a superior performance compared with other metrics:

Table 1: *Acoustic features used in the QbESOS system.*

Description	# features
Sum of auditory spectra	1
Zero-crossing rate	1
Sum of RASTA style filtering auditory spectra	1
Frame intensity	1
Frame loudness	1
Root mean square energy and log-energy	2
Energy in frequency bands 250-650 Hz and 1000-4000 Hz	2
Spectral Rolloff points at 25%, 50%, 75%, 90%	4
Spectral flux	1
Spectral entropy	1
Spectral variance	1
Spectral skewness	1
Spectral kurtosis	1
Psychoacoustical sharpness	1
Spectral harmonicity	1
Spectral flatness	1
Mel-frequency cepstral coefficients	16
MFCC filterbank	26
Line spectral pairs	8
Cepstral perceptual linear predictive coefficients	9
RASTA PLP coefficients	9
Fundamental frequency	1
Probability of voicing	1
Jitter	2
Shimmer	1
log harmonics-to-noise ratio	1
LCP formant frequencies and bandwidths	6
Formant frame intensity	1
Deltas	102
Total	204

$$\mathrm{r}(\mathbf{q}_i, \mathbf{d}_j) = \frac{\mathrm{F}(\mathbf{q}_i \cdot \mathbf{d}_j) - \|\mathbf{q}_i\|\|\mathbf{d}_j\|}{\sqrt{(\mathrm{F}\|\mathbf{q}_i^2\| - \|\mathbf{q}_i\|^2)(\mathrm{F}\|\mathbf{d}_j^2\| - \|\mathbf{d}_j\|^2)}} \tag{4}$$

where $\mathbf{q}_i \cdot \mathbf{d}_j$ denotes the dot product of $\mathbf{q}_i$ and $\mathbf{d}_j$. A mapping function is applied to $\mathrm{r}(\mathbf{q}_i, \mathbf{d}_j)$ in order to turn it into a cost function defined in [0,1]:

$$\mathrm{c}(\mathbf{q}_i, \mathbf{d}_j) = \frac{1 - \mathrm{r}(\mathbf{q}_i, \mathbf{d}_j)}{2} \tag{5}$$

Once matrix $\mathbf{M}$ is computed, the best alignment path between Q and D can be obtained following the S-DTW algorithm. First, the end of the best alignment path b^* is selected as the lowest cumulative cost among all the possible ones:

$$b^* = \underset{b \in 1,\ldots,m}{\arg\min}\, \mathrm{M}_{n,b} \tag{6}$$

After obtaining the end of the matching path, its starting point a^* is computed by backtracking the path with the lowest cost starting at b^*. In case several occurrences of query Q must be detected in document D, n-best paths can be achieved by backtracking the n end points with the lowest cost.

A score must be assigned to each detection of a query Q in a document D in order to measure how reliable that detection is. In this system, first the cumulative cost of the warping path is length-normalised [26] and, after that, z-norm is applied [27].

3. Experimental framework

The experimental framework of MediaEval 2014 Query by Example Search on Speech task (QUESST 2014) [6] was used to assess the approach proposed in this paper. This database is multilingual, since it includes speech in six different languages, namely Albanian, Romanian, Basque, Czech, non-native English, Slovak. In addition, there is a high data variability in terms of speaker, discourse and acoustic conditions, as the data includes read and spontaneous speech, broadcast speech and lectures recorded in clean and noisy scenarios. This database tries to simulate a real search application, so the queries were recorded in an isolated manner (not cut from longer recordings), and there are three different types: exact matches (T1), morphological variations (T2) and both syntactic and morphological differences (T3), which imply word reordering and word inflections [6]. Two different experiments were defined for QUESST 2014 evaluation, namely development (Dev) and evaluation (Eval); they share the same search documents but the queries differ, as summarised in Table 2.

Table 2: *Summary of QUESST 2014 database.*

		# Queries			
Partition	# Documents	Total	T1	T2	T3
Dev	12492	560	307	190	155
Eval		555	307	179	156

Two different evaluation metrics were used in this work to assess search on speech performance, namely the maximum term weighted value (MTWV)[1] and the minimum normalised cross-entropy cost $\mathrm{minC_{nxe}}$ [28], in accordance with the experimental protocol defined for QUESST 2014. It must be noted that these metrics were adopted instead of actual TWV and actual C_{nxe}, in order to ignore performance loss caused by calibration issues.

4. Experiments and results

Before presenting the experimental results, some configuration aspects of the system deserve a mention. In the gender classification approach, the features used were 19 MFCCs augmented with energy, delta and acceleration coefficients, and only voiced frames were considered. The GMMs were trained using the FA sub-corpus of Albayzin database [29], which comprises around 4 hours of speech uttered by 200 different speakers (100 of each gender). The number of mixtures of the GMMs was empirically set to 1024. With respect to the voice conversion strategy used to transform the queries, the parameters $\mathrm{f_a}$, $\mathrm{f_b}$ and k were set to 700 Hz, 3000 Hz and 0.5 according to [16], while the parameter α was set to $\pi/24$ for male-to-female conversion and $-\pi/24$ for female-to-male conversion. Given that the queries of QUESST 2014 database were recorded as isolated words, they have silence intervals before and after the actual query. These silence intervals were automatically removed using the voice activity detection approach described in [30]. With respect to the feature selection approach, and following the results reported in [15], the most relevant 130 features were used for audio representation.

The top rows of Table 3 show the performance achieved when searching the documents, regardless query and document gender, both using original and converted queries. The Table also shows the results achieved when applying the gender variability compensation approach proposed in Section 2. An increase of performance by 3% and 2% in terms of MTWV was

[1] The Spoken Term Detection (STD) 2006 Evaluation Plan, National Institute of Standards and Technology (NIST): `http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf`

obtained on Dev and Eval data, respectively. This effect was also observed in terms of $\mathrm{minC}_{\mathrm{nxe}}$. Since no strategy for dealing with query variations (types T2 and T3) was implemented, it is interesting to see the results achieved on type T1 queries only, which is shown on the bottom rows of Table 3. In this case, it can be seen that the results are significantly higher than when evaluating all types of queries, as expected. In addition, the same performance boost when applying the gender variability compensation technique is observed.

Table 3: *Dev and Eval results on QUESST 2014 experimental framework when using original and converted queries, and when combining them using the proposed gender variability compensation (VC) approach. Results are shown for all queries and for type T1 queries.*

		Dev		Eval	
	System	MTWV	$\mathrm{minC}_{\mathrm{nxe}}$	MTWV	$\mathrm{minC}_{\mathrm{nxe}}$
All	Original	0.129	0.861	0.116	0.863
	Converted	0.140	0.862	0.113	0.860
	Gender VC	**0.159**	**0.850**	**0.134**	**0.848**
	Side-info	0.128	0.862	0.116	0.863
T1	Original	0.189	0.782	0.193	0.792
	Converted	0.236	0.781	0.193	0.789
	Gender VC	**0.259**	**0.763**	**0.223**	**0.769**
	Side-info	0.188	0.782	0.194	0.793

In order to further validate the proposed approach, it was compared with the strategy proposed in [17], where side information was used during the calibration process of the search scores to try to cope with channel and language mismatch. Hence, an experiment was performed in which the log-likelihood scores obtained by the gender classifier, both for documents and queries, were used as side information for score calibration. As shown in Table 3, the proposed gender variability compensation approach outperforms this strategy, which did not succeed at improving the performance of the reference system.

Further experiments were done in order to observe whether the improvements in performance shown in Table 3 resulted from searching for queries within documents spoken by speakers of the same gender. Hence, individual experiments were done considering all the possible combinations of query and document genders (female queries - female documents, male queries - male documents, female queries - male documents, male queries - female documents). Figure 2 shows the results of those experiments, were the blue bars correspond to the MTWV achieved with the original queries (so gender matching occurs in the two experiments on the left) and the red bars represent the MTWV obtained with the converted queries (so gender matching occurs in the two experiments on the right). As shown in the Figure, MTWV is bigger when the gender of documents and queries is equal, and this effect is observed in both Dev and Eval experiments, which proves the validity of the technique presented in this work.

5. Conclusions and future work

This paper presented a strategy to compensate the acoustic variability caused by the speakers' gender in query-by-example search on speech. Experimental validation was carried out in the framework of MediaEval 2014 Query by Example Search on Speech task (QUESST 2014) and the obtained results proved the validity of the proposed approach, since it succeeded at outperforming the reference system where no gender variability compensation was performed. This strategy was also compared with a system that used the log-likelihood ratios of queries and documents, provided by a gender classification system, as side information for score calibration, and the results of the gender variability compensation approach showed to be superior in all the experimental cases.

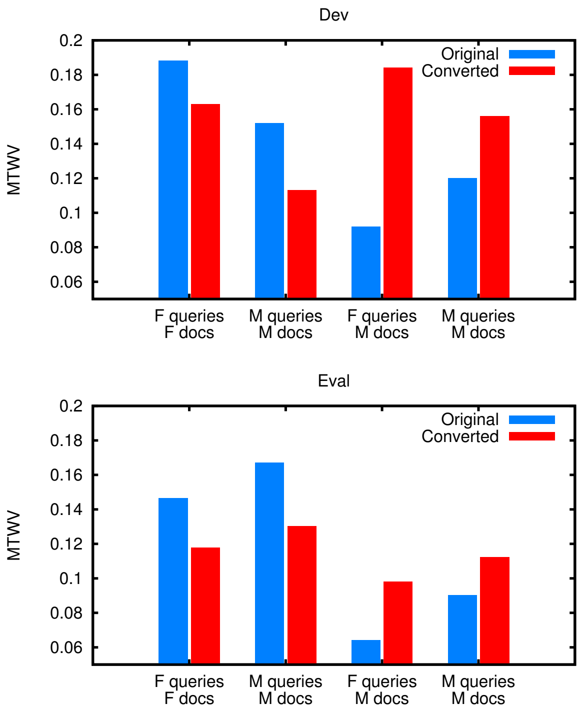

Figure 2: *MTWV obtained on Dev and Eval data when evaluating all the possible combinations of male/female (M/F) queries and documents using original and converted queries.*

The proposed strategy made use of a voice conversion approach to transform the gender of a query into its opposite. This simple technique for variability compensation enhanced the results of the reference system, suggesting that more sophisticated approaches might improve the results presented in this paper. Specifically, in future work, voice conversion will be used to transform both queries and documents into the same (or a similar) voice, since this procedure is expected to reduce the acoustic variability to a great extent, hence boosting the performance of zero-resource query-by-example search on speech strategies.

6. Acknowledgements

This research was funded by the Spanish Government (project 'TraceThem' TEC2015-65345-P), the Galician Government through the research contract GRC2014/024 (Modalidade: Grupos de Referencia Competitiva 2014) and AtlantTIC CN2012/160, the European Regional Development Fund (ERDF) and the COST Action IC1206. The authors would like to thank Carmen Magariños for providing the voice conversion software.

7. References

[1] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, no. 1, 1978.

[2] F. Metze, N. Rajput, X. Anguera, M. Davel, G. Gravier, C. V. Heerden, G. Mantena, A. Muscariello, K. Pradhallad, I. Szöke, and J. Tejedor, "The spoken web search task at MediaEval 2011," in *Proceedings of the 37th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[3] F. Metze, E. Barnard, M. Davel, C. V. Heerden, X. Anguera, G. Gravier, and N. Rajput, "The spoken web search task," in *Proceedings of the MediaEval 2012 Workshop*, 2012.

[4] X. Anguera, F. Metze, A. Buzo, I. Szöke, and L. Rodriguez-Fuentes, "The spoken web search task," in *Proceedings of the MediaEval 2013 Workshop*, 2013.

[5] J. Tejedor, D. Toledano, P. Lopez-Otero, P. Docio-Fernandez, and C. Garcia-Mateo, "Comparison of ALBAYZIN query-by-example spoken term detection 2012 and 2014 evaluations," *EURASIP Journal on Audio, Speech, and Music Processing*, 2016.

[6] X. Anguera, L. Rodriguez-Fuentes, I. Szöke, A. Buzo, and F. Metze, "Query by example search on speech at Mediaeval 2014," in *Proceedings of the MediaEval 2014 Workshop*, 2014.

[7] I. Szöke, L. Rodriguez-Fuentes, A. Buzo, X. Anguera, F. Metze, J. Proenca, M. Lojka, and X. Xiong, "Query by example search on speech at Mediaeval 2015," in *Proceedings of the MediaEval 2015 Workshop*, 2015.

[8] J. Tejedor and D. Toledano, "The ALBAYZIN 2016 search on speech evaluation plan," 2016. [Online]. Available: https://iberspeech2016.inesc-id.pt/wp-content/uploads/2016/06/EvaluationPlanSearchonSpeech.pdf

[9] T. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU*, 2009, pp. 421–426.

[10] L. Rodriguez-Fuentes, A. Varona, M. Penagarikano, G. Bordel, and M. Diez, "High-performance query-by-example spoken term detection on the SWS 2013 evaluation," in *Proceedings of the 37th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7869–7873.

[11] A. Abad, L. Rodriguez-Fuentes, M. Penagarikano, A. Varona, and G. Bordel, "On the calibration and fusion of heterogeneous spoken term detection systems," in *Proceedings of Interspeech*, 2013, pp. 20–24.

[12] M. Martinez, P. Lopez-Otero, R. Varela, A. Cardenal-Lopez, L. Docio-Fernandez, and C. Garcia-Mateo, "GTM-UVigo systems for Albayzin 2014 search on speech evaluation," in *Iberspeech 2014: VIII Jornadas en Tecnología del Habla and IV SLTech Workshop*, 2014.

[13] J. Vavrek, M. Pleva, and J. Juhár, "TUKE MediaEval 2012: spoken web search using DTW and unsupervised SVM," in *Proceedings of the MediaEval 2012 Workshop*, 2012.

[14] M. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid evaluation of speech representations for spoken term discovery," in *Proceedings of Interspeech*, 2011, pp. 821–824.

[15] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "Finding relevant features for zero-resource query-by-example search on speech," *Speech Communication*, vol. 84, pp. 24–35, 2016.

[16] C. Magariños, P. Lopez-Otero, L. Docio-Fernandez, D. Erro, E. Banga, and C. Garcia-Mateo, "Piecewise linear definition of transformation functions for speaker de-identification," in *Proceedings of First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE)*, 2016, pp. 1–5.

[17] I. Szöke, M. Skácel, L. Burget, and J. Černocký, "Coping with channel mismatch in query-by-example - BUT QUESST 2014," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5838–5842.

[18] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models." *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[19] J. Hillenbrand and M. Clark, "The role of f0 and formant frequencies in distinguishing the voices of men and women," *Attention, Perception, & Psychophysics*, vol. 71, no. 5, pp. 1150–1166, 2009.

[20] T. Zorila, D. Erro, and I. Hernaez, "Improving the quality of standard GMM-based voice conversion systems by considering physically motivated linear transformations," *Communications in Computer and Information Science (ISSN: 1865-0929)*, vol. 328, pp. 30–39, 2012.

[21] D. Erro, A. Alonso, L. Serrano, E. Navas, and I. Hernáez, "Interpretable parametric voice conversion functions based on Gaussian mixture models and constrained transformations," *Computer Speech and Language*, vol. 30, pp. 3–15, 2015.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE - the Munich versatile and fast open-source audio feature extractor," in *Proceedings of ACM Multimedia (MM)*, 2010, pp. 1459–1462.

[23] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "Phonetic unit selection for cross-lingual query-by-example spoken term detection," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, 2015, pp. 223–229.

[24] M. Müller, *Information Retrieval for Music and Motion*. Springer-Verlag, 2007.

[25] I. Szöke, M. Skácel, and L. Burget, "BUT QUESST2014 system description," in *Proceedings of the MediaEval 2014 Workshop*, 2014.

[26] A. Abad, R. Astudillo, and I. Trancoso, "The L2F spoken web search system for Mediaeval 2013," in *Proceedings of the MediaEval 2013 Workshop*, 2013.

[27] I. Szöke, L. Burget, F. Grézl, J. Černocký, and L. Ondel, "Calibration and fusion of query-by-example systems - BUT SWS 2013," in *Proceedings of the 37th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7899–7903.

[28] L. Rodriguez-Fuentes and M. Penagarikano, "MediaEval 2013 spoken web search task: system performance measures," Software Technologies Working Group, University of the Basque Country, Tech. Rep., May 2013. [Online]. Available: http://gtts.ehu.es/gtts/NT/fulltext/rodriguezmediaeval13.pdf

[29] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. Mariño, and C. Nadeu, "Albayzin speech database: design of the phonetic corpus," in *EUROSPEECH*, 1993.

[30] S. Basu, "A linked-HMM model for robust voicing and speech detection," in *Proceedings of International conference on acoustics, speech and signal processing (ICASSP)*, vol. 1, 2003, pp. 816–819.