IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Converted Mel-Cepstral Coefficients for Gender Variability Reduction in Query-by-Example Spoken Document Retrieval

Paula Lopez-Otero[1], *Laura Docio-Fernandez*[2]

[1]Universidade da Coruña - CITIC, Information Retrieval Lab
[2]Universidade de Vigo - atlanTTic Research Center, Multimedia Technology Group
paula.lopez.otero@udc.gal, ldocio@gts.uvigo.es

Abstract

Query-by-example spoken document retrieval (QbESDR) is a task that consists in retrieving those documents where a given spoken query appears. Spoken documents and queries exhibit a huge variability in terms of speaker, gender, accent or recording channel, among others. According to previous work, reducing this variability when following zero-resource QbESDR approaches, where acoustic features are used to represent the documents and queries, leads to improved performance. This work aims at reducing gender variability using voice conversion (VC) techniques. Specifically, a target gender is selected, and those documents and queries spoken by speakers of the opposite gender are converted in order to make them sound like the target gender. VC includes a resynthesis stage that can cause distortions in the resulting speech so, in order to avoid this, the use of the converted Mel-cepstral coefficients obtained from the VC system is proposed for QbESDR instead of extracting acoustic features from the converted utterances. Experiments were run on a QbESDR dataset in Basque language, and the results showed that the proposed gender variability reduction technique led to a relative improvement by 17% with respect to using the original recordings.

Index Terms: query-by-example spoken document retrieval, dynamic time warping, voice conversion, variability compensation

1. Introduction

Spoken document retrieval (SDR) consists in, given a set of spoken documents, retrieving those documents where a given query appears. The availability of technologies to perform this task is of paramount importance nowadays due to the amount of multimedia contents that are part of our everyday life. STD can be carried out using either written or spoken queries. The latter alternative, known as query-by-example SDR (QbESDR), allows a natural communication with devices while easing the access to such technologies to visually impaired users. For these reasons, this task has gained the attention of the research community, which led to the organization of evaluations in order to encourage investigation on this topic [1, 2, 3, 4, 5, 6, 7].

The most common strategies for QbESDR are based on pattern matching techniques: first a set of features is extracted from the documents and queries, and then each query-document pair is compared using an alignment algorithm usually based on dynamic time warping (DTW) [8] or any of its variants. These techniques usually rely on posteriorgram representations of the speech utterances such as phone posteriorgrams [9], which account for the probability of each phone unit in a phone decoder given a speech frame [9, 10, 11]; and Gaussian posteriorgrams, which represent the likelihood of each Gaussian in a Gaussian mixture model (GMM) given a speech frame [12, 13, 14, 15]. Gaussian posteriorgrams are computed from acoustic features extracted from the waveforms, so they can be considered a zero-resource representation for QbESDR: this is interesting since no linguistic resources are necessary to develop the system. However, the use of acoustic features for speech representation makes the system sensitive to different sources of variability such as speaker identity, gender, accent or recording channel [16].

Previous work showed that compensating the query and document variability in terms of gender leads to an improvement of QbESDR performance [16]. In those experiments, alternative queries were generated via voice conversion (VC) in order to have a female and male version of every query; then, male queries were searched within male documents and female queries within female documents. These results led to believe that transforming both queries and documents into a similar voice would reduce the acoustic variability even more. Some experiments following this research direction were presented in [14], where vocal tract length normalization (VTLN) was used to reduce the speaker-specific variation in the queries and documents: this procedure implies computing the warping factor of every recording. In this paper, a new gender variability compensation strategy is proposed that consists in, given a target gender, converting all the documents and queries of the opposite gender to that target gender so that all the resulting recordings will all be spoken by speakers of the same gender. For this purpose, a speaker-independent VC strategy [17] is used, which allows the use of the same conversion function for all the spoken utterances.

Typical VC techniques consist in a speech analysis stage followed by feature conversion and speech resynthesis in order to obtain new speech utterances with the converted speech. One of the main drawbacks of these strategies is the negative effect caused by the vocoder since it can introduce distortions that reduce the quality and intelligibility of the resulting spoken utterances [18]. Nevertheless, it would not be necessary to resynthesize the speech if the features used in the VC procedure are suitable for speech representation in the QbESDR task. In this situation, the use of Gaussian posteriorgrams for speech representation is straightforward, since it allows the modelling of the documents and queries using the converted features obtained from the VC system. Therefore, in this paper, a comparison of QbESDR results when using converted features and when extracting the same features from the converted waveforms is made in order to quantify the influence of the speech resynthesis stage.

The rest of this paper is organized as follows: Section 2 describes the proposed gender variability compensation strategy for QbESDR; the experimental framework used for validation is described in Section 3; Section 4 discusses the experimental results; and, finally, some conclusions and future work are

10.21437/IberSPEECH.2018-18

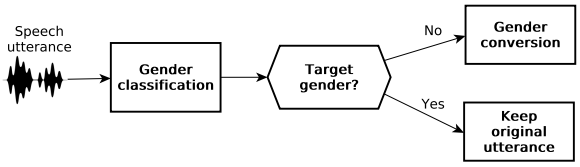

Figure 1: *Overview of the proposed approach for gender variability compensation.*

presented in Section 5.

2. Gender variability compensation in QbESDR

The gender variability compensation technique proposed in this paper is depicted in Figure 1: given a target gender, all the speech utterances that are spoken by speakers of the opposite gender are converted to the target gender via VC, while the others remain unchanged. Then, QbESDR is performed using the resulting documents and queries. In this way, all the speech utterances sound as spoken by speakers of the same gender, therefore reducing the gender variability of the recordings in the QbESDR task.

The implementation of this system requires three different tasks: gender classification, gender conversion, and search.

2.1. Gender classification

First, the gender of the documents and queries must be detected in order to find out whether a speaker belongs to the target gender or, on the contrary, the speech must be converted. In this work, a gender classifier based on the GMM log-likelihood ratio was used. Given a speech utterance to classify, its likelihoods given male and female GMMs are computed and then the log-likelihood ratio is calculated. The utterance is classified as belonging to the gender with the highest likelihood [19].

As mentioned in [16], the proposed gender classifier has shown an accuracy close to 100%, so few utterances are expected to be misclassified. Nevertheless, since the proposed approach aims at converting the gender of the utterance, the apparent gender is more relevant than the actual gender, so possible classification errors are not really important in this work.

2.2. Gender conversion

Male and female voices are different for anatomical reasons: males usually have longer vocal tracts as well as longer and heavier vocal folds than females. This derives in differences in the fundamental frequency (F0) and formant frequencies of their voice, which are generally higher for women [20]. Hence, VC techniques can be used to transform the gender of a speaker by modifying the voice characteristics of a source speaker in order to make it sound like a target speaker. In previous work on gender conversion for QbESDR, the VC technique proposed in [17] was used since it is speaker-independent, i.e. it can be used to transform any speaker into a different (undetermined) one of the opposite gender.

The gender conversion system used in this system has three stages, as most VC strategies: speech analysis, feature conversion, and speech resynthesis, as depicted in Figure 2. Three different sets of features are commonly extracted in the first stage [21]: 40 Mel-cepstral coefficients (MCEP), fundamental

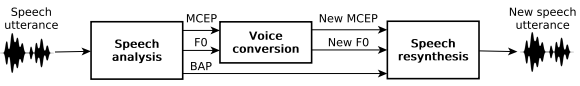

Figure 2: *Voice conversion pipeline.*

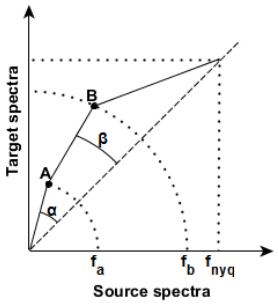

Figure 3: *Piecewise linear approximation of a FW function.*

frequency (F0) and band aperiodicity features (BAP).

The VC approach used in this work is based on frequency warping (FW) and amplitude scaling (AS), and it consists in applying an affine transformation in the cepstral domain [22]:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \quad (1)$$

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

The method proposed in [17] uses a simplified FW curve that is defined piecewise by means of three linear functions, as depicted in Figure 3: the discontinuities of the FW curve are placed at frequencies f_a and f_b; α is the angle between the 45-degree line and the first linear function; and β is the angle between the 45-degree line and the second linear function, defined as $\beta = k\alpha$ $(0 < k < 1)$. Values of α greater (less) than 0 lead to higher (lower) formant frequencies, resulting in a male-to-female (female-to-male) conversion function. This strategy is similar to vocal tract length normalization but, in this case, the same parameters are used for all the speech utterances, avoiding the need to compute a warping factor for each speaker.

Afterwards, the AS vector $\mathbf{b}$ is defined by selecting random values from a set of weighted Hanning-like bands equally spaced in the Mel-frequency scale [23] as described in [17]. Finally, the fundamental frequency is scaled proportionally to the value of α [17].

Once the MCEP and F0 features are converted (BAP features remain unchanged), speech resynthesis is performed to obtain waveforms with the converted speech. This is done using a vocoder and, depending on the goodness of the converted features and the vocoder, the resulting speech can be more or less natural and intelligible [18].

2.3. Search

The QbESDR system proposed in this work belongs to the family of pattern-matching techniques for search on speech. An overview is presented in Figure 4.

Previous work on QbESDR using gender conversion [16] relied on a large set of features plus feature selection for speech representation [24]. Nevertheless, as mentioned above, this paper aims to straightforwardly use the converted MCEP features for QbESDR and compare the performance with that achieved when extracting features from the converted waveforms. Therefore, there are two alternatives for speech representation: using

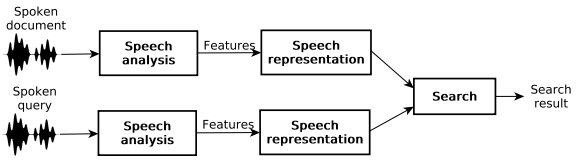

Figure 4: *Overview of the search on speech system.*

MCEP features or obtaining a more complex representation that makes use of these features. Since raw acoustic features do not usually exhibit a good performance in QbESDR, Gaussian posteriorgrams were used for this purpose: a Gaussian posteriorgram represents each frame of a spoken utterance by means of a vector of dimension G: each element of this vector is the posterior probability of each of the G Gaussians in a GMM given the frame. This representation was first proposed in [12] and used for QbESDR in [13, 14, 15], to cite some examples.

After feature extraction, given a query $\mathrm{Q} = \{\mathbf{q_1}, \ldots, \mathbf{q}_\mathrm{n}\}$ and a document $\mathrm{D} = \{\mathbf{d_1}, \ldots, \mathbf{d}_\mathrm{m}\}$ of n and m frames respectively, with vectors $\mathbf{q}_i$, $\mathbf{d}_j \in \Re^G$ and $\mathrm{n} \ll \mathrm{m}$, DTW finds the best alignment path between these two sequences. Subsequence DTW [25] (S-DTW) was used in this system, since it allows the partial alignment of a short sequence (the query) with a longer sequence (the document). The first step consists in computing a cumulative cost matrix $\mathbf{M} \in \Re^{\mathrm{n}\times\mathrm{m}}$ for a given query and document as follows:

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

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

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

In this paper, the log cosine similarity was used as the cost function as in [10] since it empirically showed a superior performance compared with other metrics:

$$cost(\mathbf{q}_i, \mathbf{d}_j) = -\log \frac{\mathbf{q}_i \cdot \mathbf{d}_j}{|\mathbf{q}_i||\mathbf{d}_j|} \tag{4}$$

This metric is normalized in order to turn it into a cost function defined in the interval [0,1]:

$$c(\mathbf{q}_i, \mathbf{d}_j) = \frac{cost(\mathbf{q}_i, \mathbf{d}_j) - cost_{min}(i)}{cost_{max}(i) - cost_{min}(i)} \tag{5}$$

where $cost_{min}(i) = \min_j cost(\mathbf{q}_i, \mathbf{d}_j)$ and $cost_{max}(i) = \max_j(\mathbf{q}_i, \mathbf{d}_j)$.

After computing $\mathbf{M}$, the S-DTW algorithm is used to find the best alignment path between Q and D. According to this algorithm, the best alignment path ends at frame b^*:

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

Then, it is possible to backtrack the whole alignment path that starts at frame a^*.

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

3. Experimental framework

The evaluation framework used in this paper is that of the QbESDR task of Albayzin 2014 search on speech evaluation. It consists of a set of spoken documents extracted from TV broadcast news in Basque language under diverse background conditions [28]. The queries were recorded in an office environment, which serves to simulate a regular user querying a retrieval system via speech. Each query includes a basic and two additional examples from different speakers; in these experiments, only the basic example is used. Two different sets of queries are included in the dataset: development (dev) queries for parameter tuning and evaluation (eval) queries to assess system performance. Table 1 summarizes some statistics of the database.

Table 1: *Summary of the experimental framework used in this paper.*

		Duration			
Data	# recordings	Total	Min	Max	# hits
Documents	1841	3 h 11 min	3.00 s	30.12 s	-
dev queries	100	2 min 51 s	1.35 s	2.29 s	772
eval queries	100	2 min 52 s	1.31 s	2.25 s	855

The evaluation metric used in this work to assess QbESDR performance is the maximum term weighted value [29], in accordance with the experimental protocol defined for Albayzin 2014 search on speech evaluation. This metric was adopted instead of actual TWV in order to ignore the performance loss caused by calibration issues.

4. Experiments and results

Before presenting the experimental results, some details of the different modules of the system described in Section 2 must be mentioned. The GMMs of the gender classification system were trained using the FA sub-corpus of Albayzin database [30], which includes around 4 hours of speech uttered by 200 different speakers (100 male, 100 female). The features used were 19 Mel-frequency cepstral coefficients (MFCCs) augmented with energy, delta and acceleration coefficients, and only voiced frames were considered. The number of mixtures of the GMMs was empirically set to 1024. The parameters f_a, f_b and k of the VC strategy were set to 700 Hz, 3000 Hz and 0.5, respectively, according to [17]. In the search stage, the silence intervals before and after the queries were automatically removed using the voice activity detection approach described in [31]. The number of Gaussians G of the GMM used for Gaussian posteriorgram computation was empirically set to 128. It must be noted that the GMM of each experiment was trained with the features extracted from its corresponding documents, so they are different for each experiment.

The first experiment aimed at comparing system performance when extracting MCEP features from the converted waveforms (*Synthesized*) and when straightforwardly using the converted MCEP features (*Converted*). As shown in Figure 5, using the converted features leads to clearly better results for dev queries. In addition, experiments were run with different values of $|\alpha|$ in order to analyze the influence of this parameter in QbESDR results. The figure shows that the best results were obtained when converting male utterances to female voices with $|\alpha| = \pi/30$. The worst performance was achieved with $|\alpha| = \pi/12$ since such a coarse conversion leads to more distorted speech according to [17]. Results with $|\alpha| = \pi/36$ are worse than those obtained with $|\alpha| = \pi/30$ because the

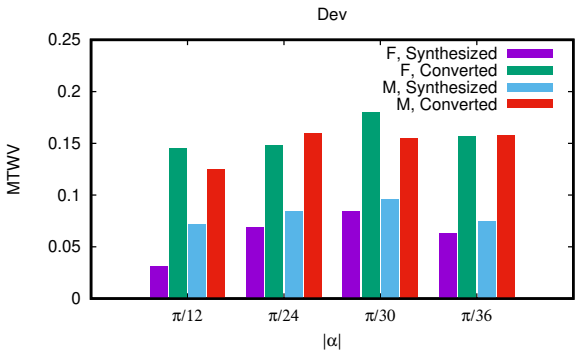

Figure 5: *Results on the dev queries when using parameters extracted from converted waveforms (Synthesized) and when using the converted parameters (Converted) for female (F) and male (M) target genders.*

conversion in this case is so subtle that it barely compensates the gender variability [17].

The experimental results on the dev queries were used to select the best target gender and $|\alpha|$ for the *Synthesized* and *Converted* systems: the best $|\alpha|$ was $\pi/30$ in both cases; the best target gender was male for the *Synthesized* experiment and female for the *Converted experiment*. Once parameter tuning was done, experiments were performed on the eval queries. Table 2 shows the results on the eval queries with the original speech (*Original*), with the *Converted* and *Synthesized* features, and when using an adapted version of the gender compensation strategy proposed in [16] (*Queries only*). The latter strategy consists in generating an opposite-gender version of each query and using the male (female) queries to search within the documents spoken by males (females). The results show that the *Converted* system achieves a relative improvement by 17% over the original speech. Performance was also evaluated for female queries with female documents (Fq-Fd), male queries with male documents (Mq-Md), female queries with male documents (Fq-Md) and male queries with female documents (Mq-Fd). The results show that the *Converted* system achieved the best performance for all the experiments. The improvement in Fq-Md and Mq-Fd experiments result from the gender variability reduction obtained with the proposed technique. Also, the results when both queries and documents were converted versions of the original ones (i.e. Mq-Md) show that the conversion is not negatively affecting the features used for speech representation. In addition, the improvement of the results when using original speech (i.e. Fq-Fd) suggest that the GMM trained with both original and converted features leads to more robust Gaussian posteriorgrams. The table shows degraded performance of the *Synthesized* system in all the experimental conditions caused by the quality of the features extracted from the resynthesized recordings. The *Queries only* system exhibits almost the same performance as the *Original* one.

Further experiments were performed in order to find out whether the variability of the recordings is reduced using the proposed technique. For this purpose, an experiment was run inspired in a state-of-art speaker verification technique: i-vectors were extracted from the *Original* and *Proposed* spoken documents, and PLDA scoring [32] of all the pairs of documents was computed. This resulted in a mean score of -5.34 and standard deviation of 11.45 for *Original* documents, and a mean score of 0.84 and standard deviation of 9.71 for *Proposed* documents.

Table 2: *Eval results for all documents and queries and for different combinations of male (M) and female (F) documents (d) and queries (d) with different systems. Conversion parameters were tuned on dev queries.*

	MTWV				
	All	Fq-Fd	Mq-Md	Fq-Md	Mq-Fd
Original	0.1659	0.2753	0.1623	0.1354	0.1969
Converted	0.1937	0.2958	0.1958	0.2013	0.2249
Synthesized	0.0835	0.2042	0.0280	0.1672	0.0272
Queries only	0.1684	0.2753	0.1623	0.1490	0.1841

This suggests that the speakers of the documents are more similar to each other when applying the proposed gender variability compensation strategy.

5. Conclusions and future work

This paper analyzed the effect of reducing gender variability via voice conversion for QbESDR. Given a target gender, all the spoken documents and queries of the opposite gender are converted in order to make them sound as spoken by the target gender. In addition, in order to alleviate the negative effects of resynthesis, the converted MCEP features were straightforwardly used to obtain Gaussian posteriorgrams of the queries and documents to perform QbESDR using a DTW-based approach. The experimental framework of the QbESDR task in Albayzin 2014 search on speech evaluation was used for validation, which consisted in a set of documents and queries in Basque language. The experiments showed a relative improvement by 17% with the proposed technique compared to the QbESDR results with the original documents and queries.

QbESDR results were analyzed according to the gender of the documents and queries, and the proposed method showed an improvement both in original and converted utterances. This suggests that a GMM trained with both original and converted features leads to more robust Gaussian posteriorgrams. In future work, the use of voice conversion for data augmentation in this scenario will be experimented.

The experimental validation showed a clear improvement when using converted features compared to extracting new features from the converted waveforms, which can be caused by the distortion introduced by the vocoder. Recent strategies for speech generation such as Wavenet have emerged in the voice conversion field, and it would be interesting to analyze the effect of different vocoders for QbESDR in the future. Also, the use of deep learning techniques for noise robust voice conversion will be assessed.

6. Acknowledgements

This work has received financial support from i) "Ministerio de Economía y Competitividad" of the Government of Spain and the European Regional Development Fund (ERDF) under the research projects TIN2015-64282-R and TEC2015-65345-P, ii) Xunta de Galicia (projects GPC ED431B 2016/035 and GRC 2014/024), and iii) Xunta de Galicia - "Consellería de Cultura, Educación e Ordenación Universitaria" and the ERDF through the 2016-2019 accreditations ED431G/01 ("Centro singular de investigación de Galicia") and ED431G/04 ("Agrupación estratéxica consolidada").

7. References

[1] 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, pp. 5165–5168.

[2] 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.

[3] 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.

[4] 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*, vol. 2016, no. 1, pp. 1–19, 2016.

[5] 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.

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

[7] J. Tejedor, D. Toledano, P. Lopez-Otero, P. Docio-Fernandez, J. Proença, F. Perdigão, F. García-Granada, E. Sanchis, A. Pompili, and A. Abad, "ALBAYZIN query-by-example spoken term detection 2016 evaluation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 2, pp. 1–25, 2018.

[8] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[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] 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.

[12] Y. Zhang and J. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU*, 2009, pp. 398–403.

[13] G. Mantena, S. Achanta, and K. Prahallad, "Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 5, pp. 944–953, 2014.

[14] M. Madhavi and H. Patil, "Vtln-warped gaussian posteriorgram for qbe-std," in *Proceedings of EUSIPCO*, 2017, pp. 593–597.

[15] ——, "Combining evidences from detection sources for query-by-example spoken term detection," in *Proceedings of APSIPA ASC*, 2017, pp. 563–568.

[16] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "Compensating gender variability in query-by-example search on speech using voice conversion," in *Proceedings of Interspeech*, 2017, pp. 2909–2913.

[17] 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.

[18] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The Voice Conversion Challenge 2018: promoting development of parallel and nonparallel methods," in *Proceedings of Odyssey*, 2018, pp. 195–202.

[19] 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.

[20] 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.

[21] F. Bahmaninezhad, C. Zhang, and J. Hansen, "Convolutional neural network based speaker de-identification," in *Proceedings of Odyssey*, 2018, pp. 255–260.

[22] 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.

[23] 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, no. 1, pp. 3–15, 2015.

[24] 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.

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

[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] J. Tejedor, D. Toledano, L. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, and G. Bordel, "The ALBAYZIN 2014 search on speech evaluation plan," 2014. [Online]. Available: http://iberspeech2014.ulpgc.es/images/EvaluationPlanSearchonSpeech.pdf

[29] J. Fiscus, J. Ajot, J. Garofolo, and G. Doddington, "Results of the 2006 spoken term detection evaluation," in *Proceedings of the ACM SIGIR Workshop "Searching Spontaneous Conversational Speech"*, 2007, pp. 51–56.

[30] 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*, vol. 1, 1993, pp. 175–178.

[31] 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.

[32] D. Garcia-Romero and C. Espy-Wilson, "Analysis of i-vector lenth normalization in speaker recognition systems," in *Proceedings of Interspeech*, 2011, pp. 249–252.