Improved External Speaker-Robust Keyword Spotting for Hearing Assistive Devices

Lopez-Espejo, Ivan; Tan, Zheng-Hua; Jensen, Jesper

Published in:
IEEE/ACM Transactions on Audio, Speech, and Language Processing

DOI (link to publication from Publisher):
10.1109/TASLP.2020.2984089

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

Improved External Speaker-Robust Keyword Spotting for Hearing Assistive Devices

Iván López-Espejo, Zheng-Hua Tan, *Senior Member, IEEE*, and Jesper Jensen

***Abstract*—For certain applications, keyword spotting (KWS) requires some degree of personalization. This is the case for KWS for hearing assistive devices, e.g., hearing aids, where only the device user should be allowed to trigger the KWS system. In this paper, we first develop a new realistic hearing aid experimental framework. Next, using this framework we show that the performance of a state-of-the-art multi-task deep learning architecture exploiting cepstral features for joint KWS and users' own-voice/external speaker detection drops significantly. To overcome this problem, we use phase difference information through GCC-PHAT (Generalized Cross-Correlation with PHAse Transform)-based coefficients along with log-spectral magnitude features. In addition, we demonstrate that working in the perceptually-motivated constant-Q transform (CQT) domain instead of in the short-time Fourier transform (STFT) domain allows for the generation of compact and coherent features which provide superior KWS performance. Our experimental results show that our CQT-based proposal achieves a relative KWS accuracy improvement of around 18% compared to using cepstral features while dramatically decreasing the number of multiplications in the multi-task architecture, which is key in the context of low-resource devices like hearing assistive devices.**

***Index Terms*—Robust keyword spotting, hearing assistive device, external speaker, constant-Q transform, generalized cross-correlation, multi-task learning.**

I. Introduction

KEYWORD spotting (KWS) is a technology concerning the identification of pre-defined keywords in utterances. KWS is in vogue in recent years thanks in part to virtual assistants such as Apple's Siri or Amazon's Alexa that are activated via voice using keywords [1]. The electronic devices on which those KWS systems run (e.g., smartphones and smart speakers) are often characterized by strict constraints in terms of computational resources [2]. This fact has encouraged further research on so-called small-footprint (i.e., low memory and low computational complexity) KWS [3]–[6].

With computational constraints in mind, attention has been progressively moving from LVCSR (Large-Vocabulary Continuous Speech Recognition)- [7] and keyword-filler HMM (Hidden Markov Model)-based KWS [8] to KWS mainly based on deep learning [3], [9], [10], as this technology can facilitate the design of compact and highly accurate models. Specifically, convolutional neural networks (CNNs) for small-footprint KWS, which tend to outperform classical deep neural networks (DNNs) [3] with far fewer parameters [11], can be considered state-of-the-art. The first attempt to use CNNs for small-footprint KWS, by Sainath and Parada [11], was recently improved by jointly integrating deep residual learning and dilated convolutions [12]. This work shows outstanding KWS accuracy results on the Google Speech Commands Dataset [13], a real-speech corpus well-suited for KWS research.

Apart from voice-based activation of virtual assistants, KWS has a number of other applications. For example, manual operation of small, body-worn devices, e.g., hearing assistive devices such as hearing aids, might be cumbersome. Changing hearing aid settings or adjusting the volume typically involves that elderly people, potentially with reduced fine motor skills, have to press small buttons on a device mounted on the ear. Even when the hearing aid can be operated by means of an app running on a smartphone, one still needs to use the hands and eyes, which can be problematic in certain situations, e.g., when driving a car. In all these scenarios, KWS can potentially provide more comfortable user interaction. A KWS system intended for hearing assistive devices should meet, at least, two important requirements:

1) A small computational and memory footprint, as hearing assistive devices are low-resource devices;
2) To only be triggered by the user of the hearing assistive device and not by any other person, i.e., an external speaker.

To the best of our knowledge, all of the small-footprint deep KWS systems above are speaker-independent, which means that any person, user or not, might trigger them. An attempt to develop personalized, that is, speaker-dependent, KWS was reported in [14]. In this work, a convolutional long short-term memory (CLSTM) model is employed to jointly perform KWS and text-dependent speaker verification. A major drawback of this multi-task approach is that KWS performance is degraded with respect to an equivalent system only dealing with the KWS task. The authors of [14] hypothesize that this drawback (which we have also observed in initial experiments) "*may be attributed to the fact that preserving speaker information may be diluting the goal of the KWS task which attempts to derive the keywords irrespective of the target speaker*" [14].

In our previous study [15], we proposed KWS for hearing assistive devices which was designed to be robust to external speakers. In particular, the small-footprint deep residual neural network of [12] was extended to jointly perform KWS and

Manuscript received month day, year; revised month day, year; accepted month day, year. Date of publication month day, year; date of current version month day, year. This work was supported, in part, by the Demant Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was xxyyzz xxyyzz.

I. López-Espejo, and Z.-H. Tan are with the Department of Electronic Systems, Aalborg University, Aalborg 9220, Denmark (e-mail: ivl@es.aau.dk; zt@es.aau.dk).

J. Jensen is with the Department of Electronic Systems, Aalborg University, Aalborg 9220, Denmark, and also with Oticon A/S, Smørum 2765, Denmark (e-mail: jje@es.aau.dk; jesj@oticon.com).

users' own-voice/external speaker detection (multi-task deep residual neural network) with a negligible increase in the number of network parameters. Thanks to exploiting *1)* cepstral features from the front and rear microphones of a hearing aid and *2)* the users' own-voice/external speaker detection, KWS accuracy results in the presence of external speakers are as good as those of an equivalent state-of-the-art KWS system [12] that does not deal with external speakers and is not constrained to any particular user. It is worth noting that this result contrasts with the KWS performance degradation observed in the aforementioned personalized (speaker-dependent) KWS system [14].

The above KWS accuracy results [15] were obtained on a speech database emulating a hearing aid as a sound capturing device. To create this database, the Google Speech Commands Dataset (GSCD) signals were filtered with impulse responses modeling acoustic channels between speakers, including both users and external speakers, and the hearing aid microphones. While this database comprises speech signals uttered by many different speakers, the impulse responses used in [15] were only measured on a single actual person wearing a hearing aid, which is referred to as "single-user" in this work. Nevertheless, it is clear that such impulse responses are user-dependent as these characterize the physical features (e.g., head size and shape) of the users. Hence, in an effort to alleviate the lack of realism of the single-user speech database used in [15], a new speech corpus with multi-user impulse responses — that is, with impulse responses measured on multiple persons wearing a hearing aid— is created in this work. We will experimentally show that, when employing this new multi-user speech database, a significant performance loss in terms of KWS accuracy can be observed for our previous multi-task deep residual neural network [15] in the presence of external speakers compared to an equivalent KWS system that does not deal with external speakers and is not constrained to any particular user [12].

Towards reducing this performance loss, we exploit the following characteristic of our hearing assistive device set-up: the relative position of the users' mouth with respect to the hearing aid front and rear microphones is virtually time-invariant and different from that of an external speaker. Thus, in this paper we explore the use of spectral magnitude and phase difference information between microphone signals — mainly intended for KWS and own-voice/external speaker detection, respectively— for our multi-task deep residual neural network for KWS robust to external speakers. In particular, for better discrimination between users' own-voice and external speakers, we deploy GCC-PHAT (Generalized Cross-Correlation with PHAse Transform)-based [16] coefficients, which are typically used to derive the time delay of arrival (TDoA) between microphones [17].

For the generation of a compact input tensor (i.e., three-dimensional matrix) integrating log-spectral magnitude and GCC-PHAT-based features, we propose the use of the perceptually-motivated constant-Q transform (CQT) [18] as an alternative to the short-time Fourier transform (STFT). Conceived for Western music processing, the CQT is characterized by geometrically-spaced filters, so at lower (higher) frequencies the frequency (time) resolution is higher. The CQT has proven to be a powerful analysis tool for different applications like audio separation [19], speaker verification [20] and speaker verification anti-spoofing [21], [22].

Furthermore, we identify three personalization dimensions which influence the microphone signals in the context of KWS for hearing assistive devices: *1)* the acoustic channel between the user's mouth and the hearing assistive device microphones, *2)* the acoustic channels between external speakers and the device microphones and *3)* the user's voice characteristics. Specifically, if a completely user-specific system is developed, strong knowledge of each dimension will be available. However, in a practical set-up, to train a user-specific system would require either the *a priori* recording of a large amount of speech data from such a specific target user or measuring impulse responses on her/him for speech data simulation, which would be time-consuming and expensive. Obviously, intermediate situations could be envisioned, where some prior knowledge is available about each dimension, e.g., gender-dependent systems. In this paper we assess the importance of dimensions *1)* and *2)* by means of variants of the multi-user database personalizing the acoustic channel between the user's mouth and the hearing assistive device microphones and/or the acoustic channels between external speakers and the user's device microphones. Our motivation for de-emphasizing the study of dimension *3)* is mainly a practical one: the GSCD, upon which we base our study, does not contain a sufficient amount of speech data from any single speaker to allow for a fair and meaningful study of this dimension. Nevertheless, in an attempt to assess the importance of dimension *3)*, at least partly, we studied the importance of users' voice characteristics from a *gender-specific* point of view (which the GSCD *does* allow for). However, these preliminary gender-dependent KWS tests[1] showed no statistically significant improvements over gender-independent approaches.

The proposed CQT-based KWS system for hearing assistive devices provides the following benefits in a multi-user scenario:

1) Relative KWS accuracy improvements of around 18% compared to using MFCCs (Mel-Frequency Cepstral Coefficients) as in our previous study [15] and 29% with respect to an equivalent system that does not deal with external speakers [12];
2) A relative KWS accuracy worsening around 1% only compared to an equivalent personalized, single-user system (i.e., trained for a specific target user in terms of acoustic channels). Indeed, we will show the superiority of a user-specific system despite its aforementioned practical disadvantages;
3) A negligible relative increase in the number of multiplications of around 0.87% with respect to the original deep residual model [12] that we extend to perform KWS robust to external speakers. This is a prominent result in

[1]To perform these tests, we manually annotated the gender of the GSCD speakers. We have made these speaker gender labels publicly available at `https://ilopezes.files.wordpress.com/2019/10/gscd_spk_gender.zip`.

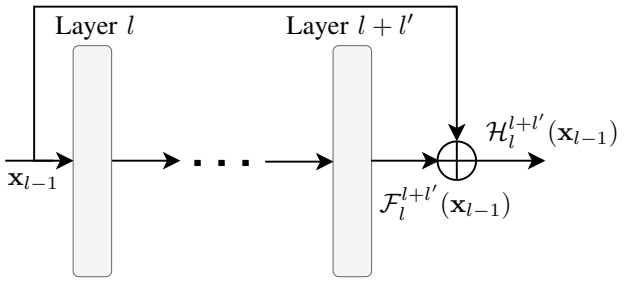

Fig. 1. General example of a residual block.

the context of low-resource devices like hearing assistive devices.

In addition, we find out that the personalization of the acoustic channel between the user's mouth and the hearing assistive device microphones is much more important to achieve a comparable performance to a fully personalized system than the personalization of the acoustic channels between external speakers and the user's device microphones. From a practical point of view, this is an important finding, since measuring impulse responses on a target user to only model the acoustic channel between her/his mouth and the hearing assistive device microphones is less cumbersome than carrying out similar measurements to model acoustic channels between external speakers and the user's device microphones.

The rest of this paper is organized as follows. In Section II, multi-task deep residual learning, considered in this work for KWS and own-voice/external speaker detection, is revisited. CQT-based features, including GCC-PHAT-based ones, for joint KWS and own-voice/external speaker detection are described in Section III. In Section IV, the new and more realistic multi-user hearing aid speech corpus and its variants are presented along with model training details. Experimental results are shown and discussed in Section V. Finally, conclusions are summarized in Section VI.

II. Multi-task Deep Residual Learning for KWS and Own-Voice/External Speaker Detection

Some fundamentals of deep residual learning are briefly reviewed in Subsection II-A. Next, in Subsection II-B, we revisit the multi-task deep residual network for joint KWS and own-voice/external speaker detection proposed in [15].

A. Deep Residual Learning

As is well-known, depth in neural networks is of importance to model arbitrarily complex functions [23]. Nevertheless, training very deep neural networks is not an easy task and, as an example of this, the extension of a well-trained model with additional layers might even lead to higher training error [24]. For better training of very deep models, He *et al.* [25] proposed residual learning.

Residual learning models can be built by concatenation of basic units called residual blocks. A diagram of a general residual block is shown in Figure 1. Let $\mathbf{x}_{l-1}$ be the input to layer l. The authors of [25] hypothesize that it might be easier to optimize the residual mapping $\mathcal{H}_l^{l+l'}(\mathbf{x}_{l-1}) = \mathcal{F}_l^{l+l'}(\mathbf{x}_{l-1}) + \mathbf{x}_{l-1}$ between layers l and $l+l'$ ($l' \in \mathbb{N}$) than the original $\mathcal{F}_l^{l+l'}(\mathbf{x}_{l-1})$ when networks are too deep. As can be seen from Figure 1, residual mapping is carried out through the so-called identity shortcut connection (which performs identity mapping) skipping $l'+1$ layers. Identity shortcut connections help to deal with the performance degradation in too deep networks.

TABLE I
Receptive field of the network in Figure 2, r_l, as a function of the convolutional layer, l.

l	1	2	3	4	5	6	7	8	9	10	11	12	13	14
r_l	3	5	7	9	13	17	21	29	37	45	61	77	93	125

Deep residual learning has been successfully applied to different tasks like noise-robust speech recognition [26] and speaker verification [27].

B. Multi-task Architecture

The multi-task deep residual neural network for joint KWS and own-voice/external speaker detection proposed in our previous work [15] (which was based on [12]) is depicted in Figure 2. This architecture is considered to be the starting point of the present study.

For speech feature extraction, audio signals captured by the front ($i = 1$) and rear ($i = 2$) microphones of the hearing assistive device are band-pass-filtered considering low and high cut-off frequencies of 20 Hz and 4 kHz, respectively [12]. Filtered signals are framed using a 30 ms Hann window with a 10 ms shift and, then, $\mathcal{Q} = 40$-dimensional MFCC vectors, $\mathbf{v}_i(t) \in \mathbb{R}^{\mathcal{Q}\times 1}$, $i = 1, 2$, are extracted from each time frame t:

$$\mathbf{v}_i(t) = [v_i(1,t), ..., v_i(q,t), ..., v_i(\mathcal{Q},t)]^\top. \quad (1)$$

In this way, a two-dimensional MFCC matrix $\mathbf{V}_i \in \mathbb{R}^{\mathcal{Q}\times T}$, that is, $\mathbf{V}_i = [\mathbf{v}_i(1), ..., \mathbf{v}_i(t), ..., \mathbf{v}_i(T)]$, $i = 1, 2$, is obtained for each microphone signal, the total number of time frames of which is T. Both MFCC matrices are stacked across the quefrency dimension to produce $\mathbf{V} \in \mathbb{R}^{2\mathcal{Q}\times T}$, namely,

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{pmatrix}. \quad (2)$$

Then, the input features to the model in Figure 2, $\tilde{\mathbf{V}} \in \mathbb{R}^{T\times 2\mathcal{Q}}$, are computed from transposing $\mathbf{V}$ and normalizing it to have zero mean and unit standard deviation. Each matrix element $\tilde{\mathbf{V}}(t,q)$, $t = 1, ..., T$, $q = 1, ..., 2\mathcal{Q}$, is defined as

$$\tilde{\mathbf{V}}(t,q) = \frac{\mathbf{V}(q,t) - \mu_{\mathbf{V}}}{\sigma_{\mathbf{V}}}, \quad (3)$$

where $\mu_{\mathbf{V}}$ and $\sigma_{\mathbf{V}}$ are the sample mean and standard deviation estimated from all the elements of $\mathbf{V}$.

As can be seen from this figure, the multi-task architecture is composed of batch normalization, average pooling and fully-connected (dense) layers as well as convolutional layers with a $\kappa \times \kappa$ kernel size, $\kappa = 3$, zero bias vectors and 45 feature maps each. This architecture uses six residual

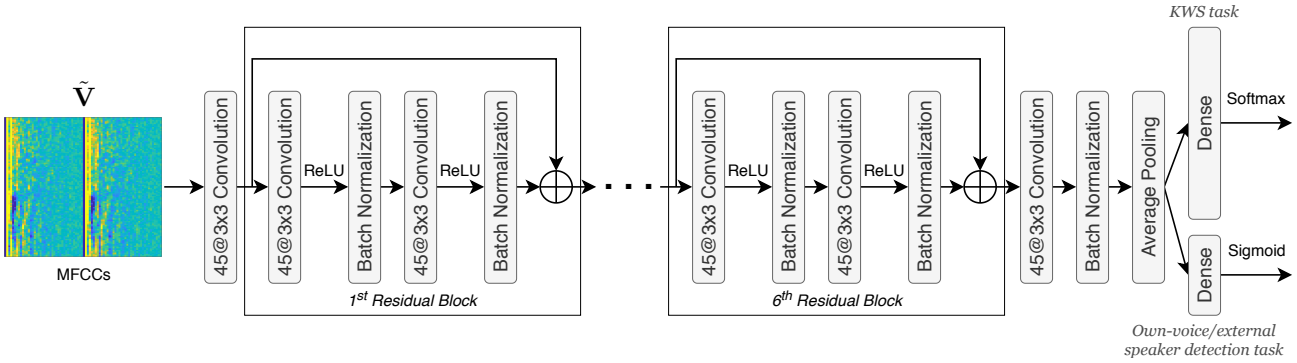

Fig. 2. Multi-task deep residual neural network for joint KWS and own-voice/external speaker detection.

blocks with identity mapping, each of them comprised of two convolutional layers followed by a rectified linear unit (ReLU) activation function. All the 14 convolutional layers in this network except the first one, $l = 1$, apply dilated convolutions [28]. The dilation rate of these convolutional layers is $(d_l, d_l) = \left(2^{\lfloor \frac{l-2}{3} \rfloor}, 2^{\lfloor \frac{l-2}{3} \rfloor}\right)$, where $l = 2, ..., 14$ and $\lfloor \cdot \rfloor$ is the floor function. Note that increasing the receptive field of the network by means of dilated convolutions helps to capture longer-range spectro-temporal dependencies of the speech signal. Table I shows the receptive field of the network, which can be easily calculated as a function of the convolutional layer l as $r_l = r_{l-1} + (\kappa - 1)d_l$, where $r_0 = d_1 = 1$. Finally, fully-connected (dense) layers with softmax and sigmoid activations are used for keyword classification and own-voice/external speaker detection, respectively.

The loss function for training this model consists of the sum of the multi-class cross-entropy loss for KWS plus the binary cross-entropy loss for own-voice/external speaker detection [29]. Specifically, let θ and C be the parameters of the model and the total number of different keywords that can be identified, respectively. Defining an additional non-keyword class, the network must solve a $(C+1)$-class classification problem plus a parallel binary classification task. Thus, $P\left(\mathcal{W}_c \middle| \tilde{\mathbf{V}}, \theta\right)$ is the posterior probability, predicted by the network, of keyword (or non-keyword) $\mathcal{W}_c$, $c = 1, ..., C+1$, given the input speech features $\tilde{\mathbf{V}}$ (whether uttered by the intended user or an external speaker). Similarly, $P\left(S_u \middle| \tilde{\mathbf{V}}, \theta\right)$ is the conditional probability, predicted by the network, that the utterance $\tilde{\mathbf{V}}$ originates from the intended user S_u. Let S_e represent an external speaker, and notice that $P\left(S_u \middle| \tilde{\mathbf{V}}, \theta\right) = 1 - P\left(S_e \middle| \tilde{\mathbf{V}}, \theta\right)$. As a result, the total cross-entropy loss function, $\mathcal{L}(\theta)$, can be expressed as

$$\mathcal{L}(\theta) = -\sum_{c=1}^{C+1} y_c \log\left(P\left(\mathcal{W}_c \middle| \tilde{\mathbf{V}}, \theta\right)\right) - \\ - y_o \log\left(P\left(S_u \middle| \tilde{\mathbf{V}}, \theta\right)\right) - (1 - y_o) \log\left(P\left(S_e \middle| \tilde{\mathbf{V}}, \theta\right)\right), \quad (4)$$

where $\{y_c; c = 1, ..., C+1\}$ and y_o are, respectively, the corresponding KWS and own-voice (binary) true labels.

At test time, keyword prediction from input features $\tilde{\mathbf{V}}$ is taken into account only if $P\left(S_u \middle| \tilde{\mathbf{V}}, \theta\right) > P_{THR}$. Rather than using $P_{THR} = 0.5$ as in [15], in this paper we follow an optimizing criterion consisting of selecting, on a model basis, the value of P_{THR} that maximizes the own-voice/external speaker detection accuracy on a validation set. In this work, accuracy is defined as the ratio between the number of correct predictions and the total number of predictions [30].

III. Constant-Q Transform-Based Features

In this section we present the log-spectral magnitude and GCC-PHAT-based features that we propose instead of MFCCs as input to the multi-task deep residual neural network depicted in Figure 2. Our goal is to design an input feature tensor (i.e., three-dimensional matrix) meeting the following criteria:

- *Compactness*: An input feature tensor with a reduced width and height still providing a competitive KWS performance is desired in order to limit the number of multiplications of the deep residual model.
- *Coherence*: To try to ease the learning of and exploit inter-feature correlations, it is desirable that the width and height axes of the different types of stacked feature matrices (in our case, log-spectral magnitude and GCC-PHAT-based matrices) correspond to the same physical units (e.g., linear frequency in hertz and time in seconds).

To meet the above criteria, we may compute features in the STFT domain. In comparison with perceptually-motivated filter banks [31], [32], the STFT is characterized by a constant frequency resolution regardless the frequency range. As is well-known, this may involve a disadvantage for a number of speech and audio signal processing applications compared to using non-linear frequency filter banks that mimic the human hearing system, e.g., [33], [34]. Since we are also interested in phase difference information, a widespread perceptually-motivated filter bank for speech signals like the Mel-scale filter bank [35], defined only for spectral magnitudes, is not a good choice. Hence, we propose the use of the perceptually-motivated CQT as a natural alternative to the STFT for the calculation of a compact and coherent input feature tensor to our deep residual model.

Note that while the CQT is closely related to the wavelet transform [36], in general, wavelet techniques are not well-suited for our purposes for computational complexity reasons. For example, wavelet transforms based on iterated filter banks require filtering the signal hundreds of times [37].

A. Constant-Q Transform

Similarly to the Fourier transform, the CQT [18], originally developed for Western music processing, is a mathematical

tool to transform a time data series to the frequency domain. However, unlike the Fourier transform, the CQT is characterized by geometrically-spaced filters.

Indeed, both the Fourier and constant-Q transforms can be interpreted as filter banks. With this in mind, let f_k ($k \in \mathbb{N}$) be the center frequency of a filter k and δf_k its bandwidth. The Q_k factor —which measures the selectivity of a filter—, is defined as [38]

$$Q_k = \frac{f_k}{\delta f_k}. \tag{5}$$

In the case of the Fourier transform, Q_k increases for increasing filter center frequencies since the bandwidth $\delta f_k = \delta f = f_s/N$, where f_s is the sampling frequency and N is the length of the analysis window, is fixed for all filters. In contrast, in the CQT, the Q_k factor is constant, i.e., $Q_k = Q \ \ \forall k$, so at lower frequencies the frequency resolution is higher, while at higher frequencies the time resolution is higher. As a result, the CQT is better in line with the human auditory system [39].

In the CQT, the center frequencies f_k can be calculated as

$$f_k = f_{min} 2^{\frac{k-1}{B}}, \tag{6}$$

where f_{min} is the center frequency of the lowest-frequency filter and B is the number of frequency bins per octave. The parameter B, establishing the time-frequency resolution trade-off, is the most important CQT parameter to be set [36]. Using (6), we can rewrite (5) in terms of B as follows:

$$Q = \frac{f_k}{f_{k+1} - f_k} = \frac{1}{2^{\frac{1}{B}} - 1}. \tag{7}$$

To let Q be constant, the length of the analysis window, N_k, changes for each filter k in such a way that $N_k \propto f_k^{-1}$. Specifically,

$$N_k = \frac{f_s}{\delta f_k} = \frac{f_s}{f_k} Q. \tag{8}$$

For convenience, let us then draw from the STFT. Let $\{x^{(t)}(n); n = 0, ..., N-1\}$ be the t-th frame of the signal $x(n)$, the STFT of $x(n)$, $X^{STFT}(k,t)$, can be expressed as

$$X^{STFT}(k,t) = \sum_{n=0}^{N-1} w(n) x^{(t)}(n) e^{-j\frac{2\pi k n}{N}}, \tag{9}$$

where $w(n)$ is the analysis window. Based on the CQT concepts introduced above, (9) is modified to define the CQT of $x(n)$, $X(k,t)$, as [18]

$$X(k,t) = \frac{1}{N_k} \sum_{n=0}^{N_k-1} w(k,n) x^{(t)}(k,n) e^{-j\frac{2\pi Q n}{N_k}}, \tag{10}$$

where, now, the length of both the signal frames $\{x^{(t)}(k,n); n = 0, ..., N_k - 1\}$ and the analysis window $w(k,n)$ changes for each filter k. In particular, we consider in this paper a Hann window:

$$w(k,n) = 0.5 - 0.5 \cos\left(\frac{2\pi n}{N_k - 1}\right), \ \ n = 0, ..., N_k - 1. \tag{11}$$

In this work, we employ the CQT implementation included in LibROSA [40] that is based on the recursive sub-sampling method described in [36].

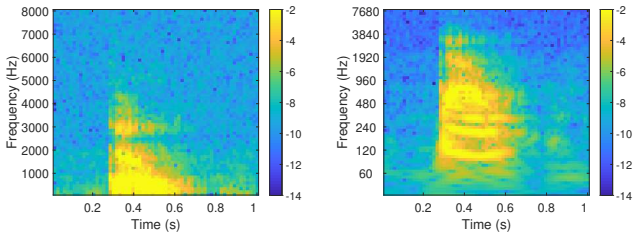

Fig. 3. Log-spectral magnitude from the STFT (left) and the CQT (right) of an utterance spoken by a user. This utterance, as captured by the front microphone of a hearing aid when worn by a user, contains the word "down" (see Section IV). Both transforms consider the same number of frequency bins.

B. Log-Spectral Magnitude and GCC-PHAT-Based Features

Below, we explain how the input feature tensor to the model is built from log-spectral magnitude and GCC-PHAT-based features calculated in the CQT domain.

1) Log-Spectral Magnitude Features: Let $x_i(n)$ be the time-domain signal captured by the i-th microphone of a hearing assistive device, where, as aforementioned, $i = 1$ and $i = 2$ refer to its front and rear microphones, respectively. The CQT of $x_i(n)$ is $X_i(k,t)$, $k = 1, ..., K$ and $t = 1, ..., T$, where K and T are the total number of frequency bins and time frames, respectively. Notice that T is large enough to cover the duration of a whole keyword. Then, the log-spectral magnitude matrices $\mathbf{X}_i \in \mathbb{R}^{T \times K}$, that is,

$$\mathbf{X}_i = \begin{pmatrix} \log\left(|X_i(1,1)|\right) & \cdots & \log\left(|X_i(K,1)|\right) \\ \vdots & \ddots & \vdots \\ \log\left(|X_i(1,T)|\right) & \cdots & \log\left(|X_i(K,T)|\right) \end{pmatrix}, \tag{12}$$

$i = 1, 2$, are jointly normalized to have zero mean and unit standard deviation and arranged into a $T \times K \times 2$ tensor to be used as input to the model.

Although in this paper we work on a two-microphone set-up, notice that the above procedure can be straightforwardly extended to an arbitrary number of microphones M to obtain a $T \times K \times M$ feature tensor.

Figure 3 shows a comparison between the log-spectral magnitudes from the STFT and the CQT of an utterance, spoken by a user, containing the word "down". This utterance, as captured by the front microphone of a hearing aid worn by a user, belongs to the hearing aid speech corpora presented in Section IV. As can be visually inspected from this figure, the CQT pays greater attention to the lower-frequency part of the spectrum, where most of the speech energy is condensed, in comparison to the STFT. It should be noticed that both transforms consider the same number of frequency bins.

2) GCC-PHAT-Based Features: A hearing assistive device like a two-microphone hearing aid is worn by a user in or behind her/his ear. As a consequence, the relative position of the user's mouth with respect to the microphones of the hearing aid is virtually time-invariant. Therefore, the phase difference between the two microphones for a particular own-voice signal should follow a recognizable pattern that is determined by a variety of factors such as the user's physical characteristics (e.g., head size and shape) and potentially influenced by the

room acoustics. Furthermore, it is reasonable to expect that, in general, such a phase difference pattern can be easily distinguished from those resulting from external speakers, whose spatial locations with respect to the microphones are necessarily different from that of a user.

Moreover, in our previous work [15], we found that the higher the similarities between the own-voice and head-related transfer functions[2] of the user in terms of MFCC Euclidean distance, the less distinguishable is an external speaker from the user. In [15], these similarities yielded a reduction in external speaker detection accuracy and, in turn, a drop in performance in terms of KWS accuracy. This is because spotting a keyword uttered by an external speaker as if it were spoken by the user is considered to be an erroneous keyword prediction.

Hence, for better discrimination between users' own-voice and external speakers, we propose the use of phase difference information through GCC-PHAT-based features in the CQT domain.

The GCC-PHAT coefficients, $G_{PHAT}(k,t)$, are defined as [16]:

$$\begin{aligned} G_{PHAT}(k,t) &= \frac{X_1(k,t)[X_2(k,t)]^*}{|X_1(k,t)[X_2(k,t)]^*|} \\ &= e^{j(\phi_1(k,t)-\phi_2(k,t))}, \\ & \qquad k=1,...,K, \quad t=1,...,T, \end{aligned} \tag{13}$$

where $|\cdot|$ denotes magnitude, $[\cdot]^*$ refers to complex conjugation, and $\phi_1(k,t)$ and $\phi_2(k,t)$ are the phases of the signals from the front and rear microphones, respectively. Then, a GCC-PHAT-based matrix $\mathbf{A} \in \mathbb{R}^{T\times K}$ is built from the angle of (13), $\measuredangle G_{PHAT}(k,t) = \phi_1(k,t) - \phi_2(k,t)$, that is,

$$\mathbf{A} = \begin{pmatrix} \measuredangle G_{PHAT}(1,1) & \cdots & \measuredangle G_{PHAT}(K,1) \\ \vdots & \ddots & \vdots \\ \measuredangle G_{PHAT}(1,T) & \cdots & \measuredangle G_{PHAT}(K,T) \end{pmatrix}. \tag{14}$$

After mean and variance normalization of $\mathbf{A}$, this matrix is stacked to the $T \times K \times 2$ log-spectral magnitude tensor described above and defined from $\mathbf{X}_i$ to create a compact and coherent $T \times K \times 3$ input feature tensor to the model.

In case of an arbitrary number of microphones, M, a total of $C_2^M = \binom{M}{2} = M(M-1)/2$ GCC-PHAT-based matrices can be calculated as in (14) from the different C_2^M pairs of microphones. In this case, the size of the feature tensor becomes $T \times K \times \left(M + C_2^M\right)$.

IV. Experimental Framework

A. Multi-user Hearing Aid Speech Corpus

A multi-user hearing aid speech database is constructed in order to train and test various variants of the proposed system. This multi-user hearing aid speech database is a generalization of the single-user hearing aid speech corpus presented in [15]. Recall that, in this paper, "single-user" and "multi-user" allude to whether impulse responses are measured, as described below, on a single person or on multiple persons wearing a hearing aid. Although the two databases have a number of features in common, the multi-user hearing aid speech database is described here in detail for the sake of completeness. The database is created from the Google Speech Commands Dataset (GSCD) [13], which is a speech corpus that contains a total of 105,829 one-second long utterances, each comprising one word among a set of 35 words. These utterances were produced by 2,618 different speakers.

[2]These concepts are carefully defined later in Subsection IV-A.

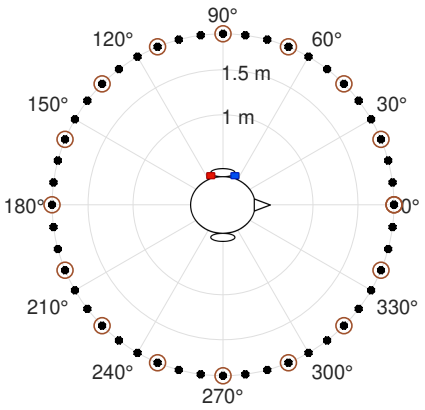

Fig. 4. Experimental set-up for transfer function measuring. Every external speaker can be located in one of the 48 equidistantly spaced points (black dots) on a circumference of 1.9 meter radius. One at a time, actual persons and mannequins wearing a two-microphone behind-the-ear hearing aid in the left ear are seated in the center of the circumference. The blue and red dots refer to the front and rear microphones, respectively, of the hearing aid. The brown circles symbolize the position of the sixteen loudspeakers.

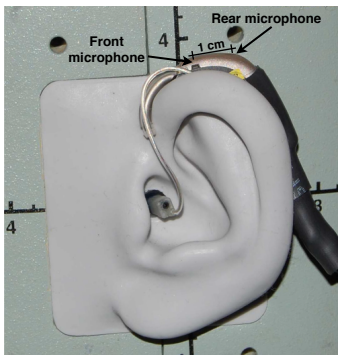

Fig. 5. Hearing aid shell mounted on the left ear of a head and torso simulator with the front and rear microphone locations annotated [41].

Figure 4 shows the experimental set-up used to generate the multi-user hearing aid speech database from the GSCD. Sixteen loudspeakers are arranged in a circular array, placed equidistantly spaced around actual female and male subjects, as well as mannequins, at eye-height in a low-reverberation listening room. Subjects and mannequins wear a two-microphone behind-the-ear hearing aid prototype in the left ear similar to the one in Figure 5 with an inter-microphone distance of 10 mm. Own-voice transfer functions (OVTFs) and head-related transfer functions (HRTFs) are measured on subjects and mannequins one at a time. An OVTF is defined as the pair of acoustic transfer functions between the mouth of the subject and the front and rear microphones of her/his left ear hearing aid. For this purpose, a close-talk microphone is placed 2 cm in front of the person's mouth and speech sentences

TABLE II
DISTRIBUTION PER DATA SET OF THE NUMBER OF USERS FOR WHOM OVTFS AND HRTFS ARE MEASURED (LEFT) AND THE NUMBER OF GSCD SPEAKERS USED TO SIMULATE USERS AND EXTERNAL SPEAKERS (RIGHT). TRANSFER FUNCTIONS ARE MEASURED ON 67 DIFFERENT SUBJECTS (FEMALES AND MALES) AND 3 DIFFERENT MANNEQUINS (ONE FEMALE AND TWO MALES). WHILE HRTFS ARE MEASURED FOR ALL OF THEM, OVTFS ARE OBTAINED ONLY FOR A SUBSET OF 29 FEMALE AND MALE SUBJECTS.

Set	No. of Users with OVTFs	No. of Users with HRTFs	No. of Speakers with a Hearing Aid	No. of External Speakers
Training	19	56	1,584	528
Validation	5	7	192	64
Test	5	7	187	63
Total	29	70	1,963	655

produced by the persons and captured by the close-talk and hearing aid microphones are used to estimate person-specific OVTFs. An HRTF is similarly defined as the pair of acoustic transfer functions between the source, i.e., loudspeaker, and the microphones of the person's or mannequin's left ear hearing aid. More in particular, a total of 48 HRTFs are measured at an angular resolution of 7.5 degrees by rotating the chair on which subjects and mannequins sit. The reader is referred to [41] for further details on this set-up.

HRTFs are measured on 67 different female and male subjects and 3 different mannequins (one female and two males), which results in a total of 70 head and torso profiles, hereinafter referred to as "users" for simplicity. OVTFs are only available for a subset of 29 female and male subjects from those 70 users [41]. From 2 to 12 HRTF takes (i.e., measurement repetitions) using either one or two distinct kinds of hearing aid prototypes similar to the one in Figure 5 are available per user. Only 1 OVTF take is available per subject.

Three different data sets are arranged in the multi-user hearing aid speech database: a training, a validation and a test. To create the training, validation and test sets, the GSCD is segmented into non-overlapping partitions of size around 80%, 10% and 10%, respectively. The 70 users are also randomly assigned to the three data sets in the same proportions. Since OVTFs are only available for 29 users, 56 users with 19 OVTFs are allocated in the training set, while the validation and test sets comprise 7 users each with 5 OVTFs. In other words, users do not overlap across sets. In addition, it is worth noticing that all mannequin HRTFs are assigned to the training set. The left part of Table II summarizes the distribution per data set of the number of users for whom OVTFs and HRTFs are measured.

For each data set, around 75% of speakers of the GSCD are chosen in a random way to simulate users, namely, subjects who wear hearing aids. Each of these speakers is randomly assigned one of the 29 users with OVTFs (see the left part of Table II). OVTFs are used to filter the corresponding GSCD signals in order to generate user speech signals. The resulting speech data constitute the so-called own-voice subsets. For each data set, the remaining 25% of speakers of the GSCD are used to simulate external speakers. Similarly, each of these speakers is randomly assigned, depending on the data set, one of the 70 users with HRTFs. On this occasion, HRTFs are utilized to filter the corresponding GSCD signals to create external speaker signals. The resulting speech data compose the external speaker subsets. Both the HRTF take[3] and the external speaker angle with respect to the simulated user (see Figure 4) are selected on an utterance basis in a uniform random manner. Finally, the distribution per data set of the number of GSCD speakers, which do not overlap across sets, used to simulate users and external speakers can be seen in the right part of Table II. Furthermore, the GSCD signals are upsampled prior to filtering them with the impulse responses, the sampling rate of which is 44.1 kHz. Filtered speech signals are then downsampled back to 16 kHz.

As in [15], models are trained to classify the following 10 different keywords: "yes", "no", "up", "down", "left", "right", "on", "off", "stop" and "go". Furthermore, the *unknown word* class, which is balanced across sets, is composed of the remaining 25 words of the GSCD. Around 10% of the utterances contains one unknown word.

B. Variants of the Multi-user Hearing Aid Speech Corpus: The SO-MH and MO-SH Corpora

To assess the impact of having personalized versus non-personalized OVTFs and/or HRTFs, we create two variants of the multi-user hearing aid speech database: the SO-MH and MO-SH corpora.

The SO-MH (Single-user OVTF-Multiple user HRTFs) corpus is just like the multi-user one except that the OVTF from only one user U is employed for all the training, validation and test sets. In a parallel manner, the MO-SH (Multiple user OVTFs-Single-user HRTFs) database only differs from the multi-user corpus in that the former utilizes HRTFs from one user only, U, for all the data sets as well. It is worth to notice that user U is also the same as that used to create the single-user hearing aid speech database of [15]. In this way, the SO-MH and MO-SH databases are subsets of the multi-user database, while the single-user database is a subset of both SO-MH and MO-SH.

C. Details on Model Training

Model training was implemented by means of Python and Keras [42] on top of TensorFlow [43]. Models were trained for a maximum of 40 epochs using early-stopping [44] with a patience of 10 epochs. For model parameter optimization, stochastic gradient descent with a momentum of 0.9 was employed using a learning rate and a learning rate decay of, respectively, 0.1 and 10^{-5}. The minibatch size was set to 64 training samples.

For regularization purposes, data augmentation was applied during training by following a similar procedure to that described in [45]. Specifically, a time shift of u ms was first applied to each utterance, where u is sampled from the uniform distribution $\mathcal{U}(-100, 100)$ on an utterance basis. Then, for each utterance, a noise segment was randomly cut with a

[3]Recall that from 2 to 12 HRTF takes, that is, measurement repetitions, are available per user.

probability of 0.8 (i.e., with a probability of 0.2, noise-based data augmentation was not considered) from one of the GSCD background noise signals. Each noise segment was scaled by a random value drawn from $\mathcal{U}(0,1)$, leading to signal-to-noise ratios (SNRs) in the wide range $[-30, 165]$ dB, before addition to the corresponding time-shifted utterance. Finally, as in [12], 30% of these distorted training data was regenerated at each epoch by re-applying the above procedure on the original training speech signals in order to increase the training data variability.

In addition, we experienced an overfitting issue leading to poor own-voice/external speaker detection performance (which, indeed, has a negative impact on KWS performance) when using CQT-based features from speech data generated from a single OVTF, i.e., the single-user database and the SO-MH database. Thus, to improve the robustness of the models, the generation procedure of training speech signals described in Subsection IV-A was modified for every database as follows. Let $h(n)$ be a transfer function between the mouth of either a user or an external speaker and either the front or rear microphone of the hearing aid. On an utterance basis, a noisy version of $h(n)$, $\tilde{h}(n)$, was used to filter the corresponding GSCD training speech signal. To obtain $\tilde{h}(n)$, an affine perturbation was applied to $h(n)$, that is,

$$\tilde{h}(n) = (1 + a_n)h(n) + b_n, \tag{15}$$

where $a_n \sim \mathcal{N}(\mu = 0, \sigma = 0.1)$ and $b_n \sim \mathcal{N}\left(\mu = 0, \sigma = 10^{-5}\right)$, and $\mathcal{N}(\mu, \sigma)$ denotes the Gaussian distribution with mean μ and standard deviation σ. We found that this sort of data augmentation procedure has a regularization effect that significantly improves the generalization ability of the models.

V. Experimental Results

Keyword spotting and own-voice/external speaker detection performance is evaluated not only on the multi-user hearing aid speech corpus but also on the single-user, SO-MH and MO-SH databases in order to assess the impact of having personalized versus non-personalized OVTFs and/or HRTFs. Our primary performance metric for both KWS and own-voice/external speaker detection is accuracy[4] with 95% confidence intervals found by means of the Student's t-distribution. Given an experiment, let μ_{acc} and σ_{acc} denote sample mean and standard deviation values, respectively, calculated from either KWS or own-voice/external speaker detection accuracy values provided by $\mathsf{n} = 10$ different networks trained with different random model initialization. Thus, confidence intervals are defined as [46]

$$\left[\mu_{acc} - t_{0.025,\mathsf{n}-1}\frac{\sigma_{acc}}{\sqrt{\mathsf{n}}},\ \mu_{acc} + t_{0.025,\mathsf{n}-1}\frac{\sigma_{acc}}{\sqrt{\mathsf{n}}}\right], \tag{16}$$

where $t_{0.025,\mathsf{n}-1} \approx 2.26$ is the 97.5$^{\text{th}}$ percentile of the Student's t-distribution with $\nu = \mathsf{n} - 1$ degrees of freedom.

A. Evaluated Techniques

Table III lists the different techniques that are evaluated along with their distinctive features. It is relevant to note that all of these techniques are two-microphone methods exploiting the front and rear microphone signals from the hearing assistive device.

As a baseline, the deep residual model for KWS with no own-voice/external speaker detection (architecture `res15`) [12] is tested. This is done to assess the performance of current KWS systems that do not deal with the potential presence of external speakers. As reviewed in Subsection II-B, the proposed multi-task architecture for KWS which is robust against external speakers, MFCC-80×1, is also evaluated. As shown in Subsection V-E, MFCC-80×1 entails a relative increase in the number of multiplications of around 105% with respect to `res15`. Then, we test MFCC-40×2 that stacks the two MFCC matrices across the depth dimension instead of across the quefrency one in order to make such a relative increase in the number of multiplications negligible (see Subsection V-E).

For tests using CQT-based features we consider a lowest-frequency filter with a center frequency of $f_{min} = 30$ Hz, 8 octaves and $B = 8$ bins per octave, so that $K = 64$ is the total number of frequency bins. With this parameter configuration we are close to spanning the entire frequency range, since $f_s = 16,000$ Hz and the upper frequency limit is $f_{max} = f_{min}2^{\frac{K}{B}} = 7,680$ Hz $\lesssim f_s/2$. In the recursive sub-sampling-based CQT implementation of [40], the analysis window shift for the highest octave that determines the amount of time frames (see [36] for further details) has to be an integer multiple of $2^{\frac{K}{B}} = 256$: it is set to 256 samples to maximize the amount of temporal information. With these choices, each microphone channel (i.e., front or rear) of every one-second long utterance with a sampling rate of 16 kHz can be represented in the CQT domain by $T = 63$ time frames with $K = 64$ frequency bins each, i.e., a total of $T \times K = 4,032$ CQT coefficients. In comparison, an MFCC-based scheme leads to 101 time frames $\times$ 40 quefrency bins $= 4,040$ MFCC coefficients[5]. These products are directly correlated to the number of multiplications in the model and, hence, to its computational complexity, so the fact that they are similar allows for a fair comparison.

The combination of our multi-task architecture along with standalone log-spectral magnitude features in the CQT domain, CQT-S, as well as with log-spectral magnitude and GCC-PHAT-based features also in the CQT domain, CQT-S+GCC, is evaluated. The results from these tests will reveal the importance of using phase difference information for improved external speaker-robust KWS. In addition, to assess the benefits of making use of the CQT instead of a transform characterized by linearly-spaced filters, equivalent tests employing the STFT, namely, STFT-S and STFT-S+GCC, are performed. For a fair comparison, $T = 63$ time frames and

[4]Recall that, in this work, accuracy is defined as the ratio between the number of correct predictions and the total number of predictions.

[5]See Table III as well as Subsection II-B to remind the MFCC extraction parameters.

TABLE III
SUMMARY OF THE DISTINCTIVE FEATURES OF THE EVALUATED TECHNIQUES.

Technique	Architecture	Training Data	Feature Domain	Type of Features	Input Dimension
Baseline	`res15` [12] (KWS only)	Own voice	Cepstrum	MFCCs	$101 \times 80 \times 1$
MFCC-80×1	Multi-task	Own & external voice	Cepstrum	MFCCs	$101 \times 80 \times 1$
MFCC-40×2	Multi-task	Own & external voice	Cepstrum	MFCCs	$101 \times 40 \times 2$
STFT-S	Multi-task	Own & external voice	STFT	Log-magnitude	$63 \times 64 \times 2$
CQT-S	Multi-task	Own & external voice	CQT	Log-magnitude	$63 \times 64 \times 2$
STFT-S+GCC	Multi-task	Own & external voice	STFT	Log-magnitude & GCC-PHAT angle	$63 \times 64 \times 3$
CQT-S+GCC	Multi-task	Own & external voice	CQT	Log-magnitude & GCC-PHAT angle	$63 \times 64 \times 3$

$K = 64$ linear frequency bins are considered by the STFT to represent each channel of every one-second long utterance.

B. Own-Voice/External Speaker Detection Results

The left part of Table IV shows the own-voice/external speaker detection accuracy results[6]. Own-voice/external speaker detection accuracy is not only measured over the whole test set (overall) but also over the own-voice and external speaker subsets of the test set separately to check possible biases towards detecting own voice or external speakers.

In the more realistic multi-user scenario (bottom row of Table IV), our previous proposal MFCC-80×1 yields poor own-voice/external speaker detection performance (around 84.26% accuracy), which is partially overcome by rearranging the input features as in MFCC-40×2. Unlike in MFCC-80×1, in MFCC-40×2, first-layer convolutions are performed over a $(101 \times 40 \times 2)$ volume in such a manner that both channels (front and rear) are merged and correlations between them are exploited in an early stage to better estimate whether the user or an external speaker is trying to trigger the KWS system. Moreover, CQT-S+GCC provides the best own-voice/external speaker detection performance so that the accuracy gain on the external speaker subset with respect to the other techniques is statistically significant. By comparing CQT-S+GCC and STFT-S+GCC against CQT-S and STFT-S we verify the convenience of exploiting GCC-PHAT-based features for differentiation between users' own-voice and external speakers. Furthermore, STFT-S is superior to CQT-S, whereas CQT-S+GCC performs better than STFT-S+GCC. This might indicate, due to the higher frequency resolution of the CQT at lower frequencies, that phase differences at lower frequencies comprise relevant information for discrimination between users' own-voice and external speakers. Besides, note that, according to performance, to work in the spectral domain is preferable to doing it in the cepstral domain.

Considering now the own-voice/external speaker detection accuracy results for the variants of the multi-user database (single-user, SO-MH and MO-SH databases) we can observe that these results are better than those from the multi-user corpus. In particular, the best results are achieved with the single-user and SO-MH corpora with no statistically significant differences between them. The main characteristic of these corpora is that they employ one OVTF only. A statistically significant degradation in own-voice/external speaker detection performance can be noticed when using multiple user OVTFs and single-user HRTFs (MO-SH database) and, even further, when employing multiple user OVTFs and HRTFs (multi-user database). It is important to bear in mind that while a single-user OVTF comprises one pair of acoustic transfer functions only, single-user HRTFs comprise 48 different pairs, that is, one per angle (see Subsection IV-A). We performed additional tests by using multiple user OVTFs and an HRTF at a single angle from a single user (modified MO-SH corpus with a single HRTF). Own-voice/external speaker detection results from these tests on a modified MO-SH corpus (not reported in this paper) are comparable to those obtained on the SO-MH corpus. Therefore, we may conclude that superior own-voice/external speaker detection performance on the single-user and SO-MH databases is due to one of the two speaker classes (i.e., the own-voice class) being fully characterized by a single OVTF that is well-learned by the neural network models.

Besides the above accuracy results, detection error trade-off (DET) curves for own-voice/external speaker detection are plotted in Figure 6. Each of these curves represents pairs of false alarm rates and false reject rates as a function of the sigmoid decision threshold, which is swept from 0 to 1. In addition, Table V summarizes estimates of the area under the curve (AUC) for the different DET curves plotted in Figure 6. Notice that the smaller the AUC, the better a system is. As can be seen from Tables IV and V, own-voice/external speaker detection accuracy results and AUC values are strongly correlated. From Table V, it is interesting to note how employing multiple OVTFs and/or multiple HRTFs yields larger AUC values due to worse own-voice/external speaker detection performance. Still, CQT-S+GCC provides the lowest AUC values for all databases but for MO-SH.

Finally, Figure 7 plots the normalized external speaker detection accuracy as a function of the angle between the external speaker and the hearing aid user. To some extent, as already discussed in [15], relationships between OVTFs and HRTFs may account for the contours of these curves. In particular, we found that OVTFs and HRTFs are more alike, in terms of Euclidean distance of input features, at angles where a relative accuracy drop can be observed. While this is particularly true for the single-user and SO-MH databases, it

[6]As explained in Subsection II-B, these results are achieved by using a sigmoid decision threshold P_{THR}, which is obtained, on a technique basis, as the threshold value maximizing the own-voice/external speaker detection accuracy on the corresponding validation set.

TABLE IV
OWN-VOICE/EXTERNAL SPEAKER DETECTION AND KWS ACCURACY RESULTS, IN PERCENTAGES, WITH 95% CONFIDENCE INTERVALS.

		Own-voice/External speaker detection			**Keyword spotting**	
		Own-voice subset	*External speaker subset*	*Overall*	*Own-voice subset*	*Overall*
Single-user database	Baseline	—	—	—	94.77 ± 0.28	72.06 ± 0.18
	MFCC-80×1	99.46 ± 0.20	97.67 ± 0.80	98.99 ± 0.18	94.99 ± 0.11	95.40 ± 0.15
	MFCC-40×2	99.93 ± 0.04	99.56 ± 0.18	99.83 ± 0.05	95.44 ± 0.24	96.49 ± 0.19
	STFT-S	99.98 ± 0.03	99.66 ± 0.18	99.89 ± 0.04	95.47 ± 0.36	96.56 ± 0.26
	CQT-S	99.97 ± 0.04	99.84 ± 0.14	99.93 ± 0.03	95.57 ± 0.27	96.67 ± 0.21
	STFT-S+GCC	99.96 ± 0.03	99.82 ± 0.08	99.93 ± 0.02	95.60 ± 0.27	96.68 ± 0.20
	CQT-S+GCC	99.99 ± 0.42	99.85 ± 0.10	99.95 ± 0.03	95.64 ± 0.20	96.74 ± 0.15
Single-user OVTF-Multiple user HRTFs (SO-MH) database	Baseline	—	—	—	94.97 ± 0.28	71.49 ± 0.19
	MFCC-80×1	99.37 ± 0.43	97.22 ± 0.64	98.79 ± 0.23	95.19 ± 0.28	95.38 ± 0.30
	MFCC-40×2	99.82 ± 0.13	99.59 ± 0.18	99.76 ± 0.09	95.71 ± 0.29	96.68 ± 0.26
	STFT-S	99.93 ± 0.09	99.29 ± 0.67	99.76 ± 0.17	95.90 ± 0.16	96.80 ± 0.13
	CQT-S	99.96 ± 0.02	99.59 ± 0.26	99.86 ± 0.06	95.88 ± 0.31	96.88 ± 0.25
	STFT-S+GCC	99.95 ± 0.03	99.86 ± 0.16	99.93 ± 0.04	95.77 ± 0.24	96.86 ± 0.16
	CQT-S+GCC	99.96 ± 0.06	99.92 ± 0.09	99.95 ± 0.05	95.96 ± 0.16	97.02 ± 0.13
Multiple user OVTFs-Single-user HRTFs (MO-SH) database	Baseline	—	—	—	94.00 ± 0.38	73.70 ± 0.28
	MFCC-80×1	96.38 ± 1.02	59.38 ± 6.48	87.87 ± 0.79	93.30 ± 0.42	83.79 ± 0.87
	MFCC-40×2	98.61 ± 0.83	91.72 ± 3.64	97.03 ± 0.61	94.31 ± 0.29	92.79 ± 0.59
	STFT-S	99.88 ± 0.06	98.09 ± 0.44	99.47 ± 0.10	94.59 ± 0.41	95.34 ± 0.37
	CQT-S	99.56 ± 0.24	97.68 ± 0.64	99.13 ± 0.15	94.53 ± 0.31	94.97 ± 0.28
	STFT-S+GCC	99.98 ± 0.03	98.57 ± 0.83	99.65 ± 0.19	94.67 ± 0.34	95.57 ± 0.18
	CQT-S+GCC	99.97 ± 0.03	98.90 ± 0.34	99.72 ± 0.07	95.25 ± 0.13	96.08 ± 0.15
Multi-user database	Baseline	—	—	—	93.81 ± 0.27	73.88 ± 0.23
	MFCC-80×1	92.64 ± 1.39	55.36 ± 4.43	84.26 ± 0.45	93.27 ± 0.30	80.45 ± 0.55
	MFCC-40×2	97.03 ± 1.81	87.18 ± 2.06	94.81 ± 1.20	94.32 ± 0.21	90.78 ± 1.16
	STFT-S	98.60 ± 0.95	95.03 ± 1.10	97.80 ± 0.53	94.30 ± 0.34	93.59 ± 0.64
	CQT-S	98.44 ± 0.87	92.12 ± 2.39	97.02 ± 0.44	94.60 ± 0.31	93.19 ± 0.52
	STFT-S+GCC	98.61 ± 1.30	96.40 ± 1.21	98.11 ± 0.93	94.23 ± 0.57	93.77 ± 0.99
	CQT-S+GCC	99.49 ± 0.47	98.67 ± 0.36	99.31 ± 0.33	94.81 ± 0.26	95.34 ± 0.32

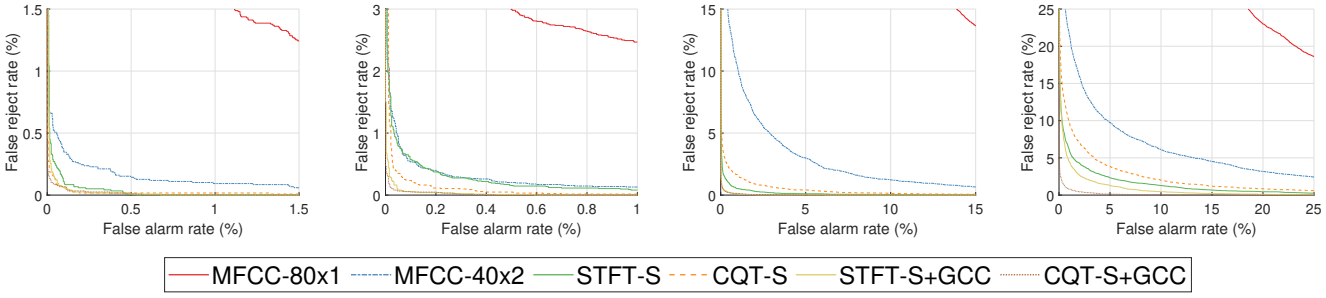

Fig. 6. Detection error trade-off curves for own-voice/external speaker detection. Different plots correspond to different databases. From left to right: single-user, SO-MH, MO-SH and multi-user databases.

TABLE V
ESTIMATION OF THE AREA UNDER THE DETECTION ERROR TRADE-OFF CURVES PLOTTED IN FIGURE 6. THE SMALLER THE AREA UNDER THE CURVE, THE BETTER A SYSTEM IS. BEST RESULTS FOR EACH DATABASE ARE MARKED IN BOLD.

Technique/Database	**Single-user**	**SO-MH**	**MO-SH**	**Multi-user**
MFCC-80×1	7.87	34.04	652.07	1,286.15
MFCC-40×2	0.37	0.64	57.69	220.78
STFT-S	0.06	0.51	2.68	45.42
CQT-S	0.04	0.16	7.66	75.79
STFT-S+GCC	0.04	0.03	**0.49**	24.88
CQT-S+GCC	**0.02**	**0.02**	0.67	**3.87**

is not so clear when exploiting MFCC features under the MO-SH and multi-user corpora. The latter may be as a result of the more complex frameworks where the own-voice and external speaker classes are characterized by multiple acoustic transfer functions each, which might have a regularization effect. In this way, the neural network models might be able to learn relevant features for discrimination between the user's own-voice and an external speaker rather than fitting to a particular pair of acoustic transfer functions. As can be seen from Figure 7, all of the evaluated techniques but our previous proposal MFCC-80×1 perform very good external speaker detection on the single-user and SO-MH corpora. External speaker detection performance of CQT-S+GCC stands out in the multi-user scenario in comparison with the other techniques, e.g., at the shadow side, namely, between 240° and 300°. Moreover, the drop in performance of CQT-S+GCC at around 60° might

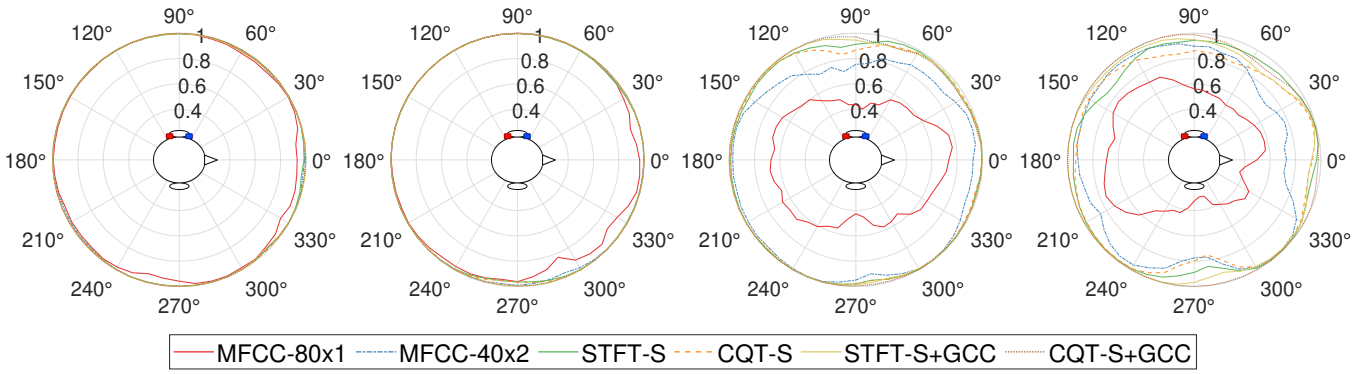

Fig. 7. Normalized external speaker detection accuracy as a function of the angle between the external speaker and the hearing aid user. The head of the users is centered in the origin and faces towards 0°. Different plots correspond to different databases. From left to right: single-user, SO-MH, MO-SH and multi-user databases.

be attributed to similar own-voice and external speaker phase differences.

Since having a computational and memory footprint as small as possible is important for hearing assistive devices, we perform equivalent technique evaluations by using an even lighter multi-task architecture based on `res15-narrow` [12]. The difference between `res15-narrow` and `res15` is that the former employs convolutional layers with 19 feature maps each instead of 45. The left part of Table VI reports the corresponding own-voice/external speaker detection accuracy results for the multi-user scenario. While the above discussion on own-voice/external speaker detection performance holds true also in this case, average accuracy results are, in general, lower when utilizing the multi-task architecture based on `res15-narrow` rather than on `res15`.

C. Keyword Spotting Results

KWS accuracy results are shown in the right part of Table IV. Apart from results on the overall test set, KWS accuracies on the own-voice subset are also presented in order to assess the degradation owing to the presence of external speakers. Thus, own-voice/external speaker detection is taken into consideration in overall KWS accuracy computation so that correct predictions are made in the following cases:

1) A user utters a keyword: The KWS system detects user's own-voice and the correct keyword.
2) A user utters a non-keyword or an external speaker utters either a keyword or a non-keyword: The KWS system detects user's own-voice and a non-keyword, or it spots an external speaker and either a keyword or a non-keyword.

From Table IV we can see that Baseline KWS accuracies are substantially higher on the own-voice subsets of the different databases than in the presence of external speakers (overall test sets). For example, in the multi-user scenario, Baseline KWS accuracy on the own-voice subset, which is around 93.81%, drops to 73.88% in the presence of external speakers. According to Table IV, the rest of the evaluated techniques tends to close this gap through the integration of own-voice/external speaker detection. Bearing in mind that KWS accuracy on the own-voice subsets is in the approximate range of 93% to 95% for all techniques, the reader can observe, as expected, a strong correlation between own-voice/external speaker detection and overall KWS accuracy. Thus, CQT-S+GCC is the best method also in terms of overall KWS accuracy and in a statistically significant manner for the multi-user and MO-SH databases. In the more realistic multi-user scenario, this method (∼95.34% acc.) achieves overall KWS accuracy relative improvements of around 18% compared to our previous proposal MFCC-80×1 (∼80.45% acc.) and 29% with respect to Baseline (∼73.88% acc.).

As presented in Section I, one of the first attempts to develop personalized (speaker-dependent) KWS [14], through joint KWS and text-dependent speaker verification, suffers from a major drawback: KWS performance is degraded with respect to an equivalent system only dealing with the KWS task. On the contrary, regardless the database, there are no statistically significant differences between the own-voice subset KWS accuracies from MFCC-80×1 and Baseline (even though own-voice detection on the own-voice subsets is not flawless), which uses the same input features as MFCC-80×1. Therefore, we can state that the above major drawback is solved by our multi-task architecture. This may be attributed to the fact that we exploit two different kinds of information for KWS and personalization (own-voice/external speaker detection), i.e., spectral and spatial information, respectively. By contrast, in [14], both KWS and personalization, which is carried out through tackling a more difficult task (i.e., speaker verification), may exhibit a certain degree of interference as they rely on the same set of spectral features.

Except for the fact that the single-user database comprises a number of different speakers, results on this database may be considered an upper bound[7] for the performance of a fully personalized KWS system, that is, a system that is intended for a specific target user. Notice that CQT-S+GCC yields an overall KWS accuracy relative worsening around 1% only between the single- (∼96.74% acc.) and the multi-user (∼95.34% acc.) scenario. Thus, apparently, CQT-S+GCC helps closing the gap between a non-personalized KWS system which is robust to external speakers and a fully personalized KWS

[7]For instance, remember that the OVTF is time-invariant and does not take into account different factors that can alter it such as the room acoustics.

TABLE VI
MULTI-USER OWN-VOICE/EXTERNAL SPEAKER DETECTION AND KWS ACCURACY RESULTS, IN PERCENTAGES, WITH 95% CONFIDENCE INTERVALS. TECHNIQUE NAMES WITH SUFFIX -N REFER TO THE USE OF A LIGHTER MULTI-TASK ARCHITECTURE BASED ON RES15-NARROW [12].

		Own-voice/External speaker detection			**Keyword spotting**	
		Own-voice subset	*External speaker subset*	*Overall*	*Own-voice subset*	*Overall*
Multi-user database	Baseline-n	—	—	—	92.53 ± 0.47	73.15 ± 0.31
	MFCC-80×1-n	90.83 ± 2.40	58.11 ± 8.48	83.47 ± 0.91	92.45 ± 0.31	79.13 ± 0.88
	MFCC-40×2-n	97.25 ± 2.02	84.24 ± 6.05	94.33 ± 1.17	92.69 ± 0.30	89.12 ± 1.22
	STFT-S-n	99.01 ± 0.59	92.46 ± 1.81	97.53 ± 0.22	92.93 ± 0.65	92.39 ± 0.57
	CQT-S-n	97.67 ± 1.64	91.12 ± 2.74	96.20 ± 1.04	92.96 ± 0.70	91.20 ± 1.19
	STFT-S+GCC-n	98.44 ± 1.41	94.26 ± 2.33	97.50 ± 0.79	92.45 ± 0.40	91.87 ± 0.86
	CQT-S+GCC-n	98.92 ± 0.86	97.94 ± 0.69	98.70 ± 0.59	93.02 ± 0.67	93.38 ± 0.67

system. This is a remarkable result, especially considering the practical disadvantages of a fully personalized KWS system as discussed in Section I. Moreover, when comparing KWS accuracy results from the SO-MH and MO-SH databases, we see that having a personalized OVTF is much more important than having personalized HRTFs. In fact, KWS accuracy results from the single-user and SO-MH corpora are very similar. Thus, we can conclude that we may only need personalized OVTFs to achieve a comparable performance to a fully personalized system. As already mentioned in Section I, this is an important finding from a practical point of view, as measuring OVTFs on a specific target user is less cumbersome than measuring impulse responses to model acoustic channels between external speakers and the user's device microphones, i.e., HRTFs.

Finally, note that the right part of Table VI reports the KWS accuracy results for the multi-user scenario when employing the lighter multi-task architecture based on `res15-narrow`. KWS accuracy trends are similar to those from using the multi-task architecture based on `res15`, whereas accuracy performance is generally better in the latter case in a significant manner (see Table IV for comparison).

D. Streaming Keyword Spotting

Real-life application of KWS generally involves that KWS systems have to process a continuous stream of audio data where the delimitation of the spoken words is unknown. Furthermore, it can reasonably be expected that, most of the time, KWS systems will hear other things rather than keywords. Hence, to figure out how our CQT-based proposal performs under these circumstances, we carry out streaming KWS evaluations in this subsection.

We train the KWS systems to recognize an additional class consisting of silence/background noise. As above, all the classes are approximately balanced for training. On the other hand, we generate a test audio stream by concatenation of test speech and silence/background noise segments. The test audio stream is comprised of around 400 keywords of each of the 10 considered types and near 7,000 non-keywords uttered by both users and external speakers. All types of words are randomly mixed along the audio stream, the total duration of which is, approximately, 3 hours and 46 minutes. For testing, we employ a one-second long sliding window with a hop of 250 ms. For streaming KWS performance evaluation, we process the sequence of modified posteriors $\tilde{P}\left(\mathcal{W}_c \middle| \tilde{\mathbf{V}}, \theta\right) = P\left(\mathcal{W}_c \middle| \tilde{\mathbf{V}}, \theta\right) \delta_u$ as in [47], where

$$\delta_u = \begin{cases} 1 & \text{if } P\left(S_u \middle| \tilde{\mathbf{V}}, \theta\right) > P_{THR}, \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Since accuracy can be a misleading metric for data sets with rather imbalanced classes, streaming KWS performance is measured by means of DET, precision/recall and F-score curves, which are computed for each of the keywords and, then, averaged across them. Figure 8 plots these average curves for the more realistic multi-user scenario and the most relevant evaluated techniques (and Baseline). Note that the larger the area under the precision/recall and F-score curves, the better a system is. Thus, CQT-S+GCC is the best performing method.

E. Computational Complexity

1) CQT versus STFT: Although the recursive sub-sampling-based CQT of [36] is a computationally efficient approach, it involves additional non-negligible computational load in comparison with the STFT. For instance, this CQT approach entails a number of low-pass filterings that is proportional to the number of octaves and around twice the fast Fourier transform computations compared to the STFT [36].

The computational burden of LibROSA's [40] CQT and STFT implementations employed in this work was evaluated on an Intel Xeon E5-2680 CPU with a clock frequency of 2.4 GHz. Processing each microphone channel for a single one-second long utterance takes 26.19 ms ± 0.112 and 1.61 ms ± 0.001 for the CQT and the STFT, respectively. These execution times were estimated over a set of around 10,000 utterances.

Similarly, extracting MFCCs from each microphone channel of a one-second long utterance takes 117.55 ms ± 1.027. In addition, one forward pass of the multi-task architecture takes 66.84 ms ± 0.55 for CQT-S+GCC and STFT-S+GCC and 204.46 ms ± 0.90 for MFCC-80×1. In summary, despite calculating the CQT is more computationally expensive than calculating the STFT, our CQT-based proposal is around 3.7 times faster than our previous system MFCC-80×1. Furthermore, computing MFCCs, which are used by some small-footprint KWS systems (e.g., [12]), is more computationally expensive than computing the CQT.

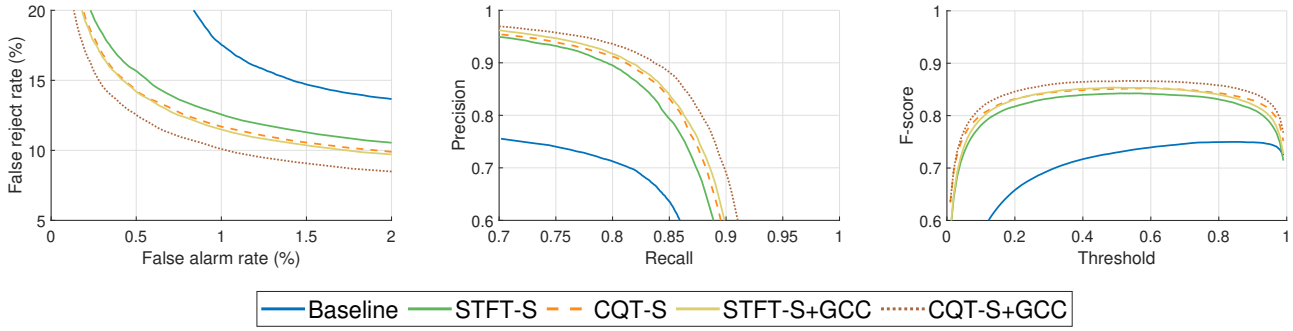

Fig. 8. Streaming KWS performance in the multi-user scenario. From left to right: detection error trade-off, precision/recall and F-score curves.

TABLE VII
NUMBER OF PARAMETERS, BROKEN DOWN BY LAYERS, OF THE MULTI-TASK ARCHITECTURE FOR JOINT KWS AND OWN-VOICE/EXTERNAL SPEAKER DETECTION. THE TOTAL NUMBER OF PARAMETERS DEPENDS ON THE DEPTH OF THE INPUT FEATURE TENSOR, D.

Layer(s)	No. of Parameters
Shallowest Conv.	$P_1 = 405 \times D$
Residual Block ($\times 6$)	$P_2 = 219,780$
Deepest Conv.	$P_3 = 18,225$
Batch Norm.	$P_4 = 90$
Avg. Pooling	$P_5 = 0$
Softmax	$P_6 = 506$
Sigmoid	$P_7 = 46$
Total	$405 \times D + 238,647$

2) Number of Parameters: Table VII reports the number of parameters, broken down by layers, of our multi-task architecture for joint KWS and own-voice/external speaker detection. The total number of parameters depends on the depth of the input feature tensor, D. Based on Table VII, the left part of Table VIII presents the number of parameters of the different techniques evaluated in this work and `res15` [12]. From Table VIII, relative increases in the number of parameters of the evaluated techniques with respect to the speaker-independent `res15` can be esteemed negligible.

TABLE VIII
NUMBER OF PARAMETERS (LEFT) AND MULTIPLICATIONS (RIGHT) OF THE EVALUATED TECHNIQUES AND RES15 ALONG WITH RELATIVE INCREASES, IN PERCENTAGES, WITH RESPECT TO RES15.

Technique	No. of Param.	Relative Inc. (%)	No. of Multip.	Relative Inc. (%)
`res15` [12]	239,006	—	897,237,540	—
Baseline	239,006	0	1,841,697,540	105.26
MFCC-80×1	239,052	0.02	1,841,697,585	105.26
MFCC-40×2	239,457	0.19	898,761,195	0.17
STFT-S	239,457	0.19	903,539,295	0.70
CQT-S	239,457	0.19	903,539,295	0.70
STFT-S+GCC	239,862	0.36	905,071,005	0.87
CQT-S+GCC	239,862	0.36	905,071,005	0.87

3) Number of Multiplications: Let H and W be the height and width of the input feature tensor, respectively, and let $F = 45$ be the number of feature maps. Moreover, recall that C is the number of different keywords that can be identified. Again based on Table VII, it can be shown that the number of multiplications in our multi-task architecture can be approximated by $(H-2)\times(W-2)\times(P_1+P_2+P_3+P_4)+F+(P_6-(C+1))+(P_7-1) = (H-2)\times(W-2)\times(405\times D+238,095)+585$. The right part of Table VIII reports the number of multiplications of the evaluated techniques and `res15`. We can see that, in contrast to our previous proposal MFCC-80×1 (and Baseline), arranging the input features to exploit the depth dimension makes the relative increases in the number of multiplications with respect to `res15` negligible.

TABLE IX
NUMBER OF PARAMETERS (LEFT) AND MULTIPLICATIONS (RIGHT) OF CQT-S+GCC−N AND REFERENCE TECHNIQUES ALONG WITH RELATIVE INCREASES, IN PERCENTAGES, WITH RESPECT TO RES15−NARROW.

Technique	No. of Param.	Relative Inc. (%)	No. of Multip.	Relative Inc. (%)
`res15-narrow` [12]	43,122	—	161,397,552	—
Baseline−n	43,122	0	331,289,472	105.26
CQT-S+GCC−n	43,484	0.84	163,549,055	1.33
CQT-S+GCC	239,862	456.24	905,071,005	460.77

Similarly, Table IX reports the number of parameters and multiplications of CQT-S+GCC−n and reference techniques. The relative increase in the number of both parameters and multiplications of CQT-S+GCC−n with respect to `res15-narrow` is minor in comparison with that of CQT-S+GCC. In return, as shown above, CQT-S+GCC is statistically significantly superior to CQT-S+GCC−n in terms of KWS accuracy.

VI. CONCLUSIONS

In this paper we have carried out a study on external speaker-robust keyword spotting for hearing assistive devices. Initially, we built a multi-user hearing aid experimental framework that is more realistic than the single-user one proposed in our previous research. Under this new framework, we have observed that the KWS performance of our multi-task architecture for joint KWS and own-voice/external speaker detection exploiting MFCC features drops substantially.

To strengthen our KWS system against external speakers, we have explored the use of phase difference information

through GCC-PHAT-based coefficients. We have demonstrated that their use along with log-spectral magnitude features provides a significantly improved KWS performance in the presence of external speakers as well as significant gains are achieved when working in the perceptually-motivated CQT domain with respect to the STFT domain. We hypothesized that the latter is due to phase differences at lower frequencies comprising relevant information for discrimination between users' own-voice and external speakers, since the CQT has a higher frequency resolution at lower frequencies in comparison with the STFT. In turn, such a good performance has been achieved while dramatically decreasing the number of multiplications with respect to our previous MFCC-based proposal. This is an important result, as hearing assistive devices, like hearing aids, are low-resource devices.

We can conclude that our findings help for closing the gap between a non-personalized KWS system robust to external speakers ready to be used by any user with no tuning and a fully personalized KWS system, which exhibits serious practical disadvantages.

Finally, although the GSCD, upon which our experimental framework was based, was recorded in real-life conditions (i.e., including noisy conditions), future work includes a systematic study on noise robustness of our KWS system.

REFERENCES

[1] M. Hoy, "Alexa, Siri, Cortana, and more: An introduction to voice assistants," *Medical Reference Services Quarterly*, vol. 37, pp. 81–88, 01 2018.

[2] S. O. Arik, M. Kliegl, R. Child, J. Hestness, A. Gibiansky, C. Fougner, R. Prenger, and A. Coates, "Convolutional recurrent neural networks for small-footprint keyword spotting," in *Proceedings of INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association, August 20-24, Stockholm, Sweden*, pp. 1606–1610, 2017.

[3] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *Proceedings of ICASSP 2014 – 39th IEEE International Conference on Acoustics, Speech and Signal Processing, May 4-9, Florence, Italy*, pp. 4087–4091, 2014.

[4] M. Sun, A. Raju, G. Tucker, S. Panchapagesan, G. Fu, A. Mandal, S. Matsoukas, N. Strom, and S. Vitaladevuni, "Max-pooling loss training of long short-term memory networks for small-footprint keyword spotting," in *Proceedings of SLT 2016 – IEEE Spoken Language Technology Workshop, December 13-16, San Diego, USA*, pp. 474–480, 2016.

[5] Y. He, R. Prabhavalkar, K. Rao, W. Li, A. Bakhtin, and I. McGraw, "Streaming small-footprint keyword spotting using sequence-to-sequence models," in *Proceedings of ASRU 2017 – IEEE Automatic Speech Recognition and Understanding Workshop, December 16-20, Okinawa, Japan*, pp. 474–481, 2017.

[6] X. Wang, S. Sun, C. Shan, J. Hou, L. Xie, S. Li, and X. Lei, "Adversarial examples for improving end-to-end attention-based small-footprint keyword spotting," in *Proceedings of ICASSP 2019 – 44th IEEE International Conference on Acoustics, Speech and Signal Processing, May 12-17, Brighton, UK*, pp. 6366–6370, 2019.

[7] D. R. H. Miller, M. Kleber, C.-L. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *Proceedings of INTERSPEECH 2007 – 8th Annual Conference of the International Speech Communication Association, August 27-31, Antwerp, Belgium*, pp. 314–317, 2007.

[8] R. Rose and D. Paul, "A hidden Markov model based keyword recognition system," in *Proceedings of ICASSP 1990 – 15th IEEE International Conference on Acoustics, Speech and Signal Processing, April 3-6, Albuquerque, USA*, pp. 129–132, 1990.

[9] R. Prabhavalkar, R. Alvarez, C. Parada, P. Nakkiran, and T. N. Sainath, "Automatic gain control and multi-style training for robust small-footprint keyword spotting with deep neural networks," in *Proceedings of ICASSP 2015 – 40th IEEE International Conference on Acoustics, Speech and Signal Processing, April 19-24, Brisbane, Australia*, pp. 4704–4708, 2015.

[10] R. Alvarez and H.-J. Park, "End-to-end streaming keyword spotting," in *Proceedings of ICASSP 2019 – 44th IEEE International Conference on Acoustics, Speech and Signal Processing, May 12-17, Brighton, UK*, pp. 6336–6340, 2019.

[11] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Proceedings of INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6-10, Dresden, Germany*, pp. 1478–1482, 2015.

[12] R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," in *Proceedings of ICASSP 2018 – 43rd IEEE International Conference on Acoustics, Speech and Signal Processing, April 15-20, Calgary, Canada*, pp. 5484–5488, 2018.

[13] P. Warden, "Speech Commands: A dataset for limited-vocabulary speech recognition," *arXiv:1804.03209v1*, 2018.

[14] R. Kumar, V. Yeruva, and S. Ganapathy, "On convolutional LSTM modeling for joint wake-word detection and text dependent speaker verification," in *Proceedings of INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India*, pp. 1121–1125, 2018.

[15] I. López-Espejo, Z.-H. Tan, and J. Jensen, "Keyword spotting for hearing assistive devices robust to external speakers," in *Proceedings of INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association, September 15-19, Graz, Austria*, 2019.

[16] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, 1976.

[17] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust TDOA estimation based on time-frequency masking and deep neural networks," in *Proceedings of INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India*, pp. 322–326, 2018.

[18] J. C. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, pp. 425–, 01 1991.

[19] R. Jaiswal, D. Fitzgerald, E. Coyle, and S. Rickard, "Towards shifted NMF for improved monaural separation," in *Proceedings of ISSC 2013 – 24th IET Irish Signals and Systems Conference, Dublin, Ireland*, 2013.

[20] H. Delgado, M. Todisco, M. Sahidullah, A. K. Sarkar, N. Evans, T. Kinnunen, and Z.-H. Tan, "Further optimisations of constant Q cepstral processing for integrated utterance and text-dependent speaker verification," in *Proceedings of SLT 2016 – IEEE Spoken Language Technology Workshop, December 13-16, San Diego, USA*, pp. 179–185, 2016.

[21] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proceedings of Odyssey 2016 – The Speaker and Language Recognition Workshop Odyssey, June 21-24, Bilbao, Spain*, pp. 283–290, 2016.

[22] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[23] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, pp. 1–127, 2009.

[24] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proceedings of CVPR 2015 – Conference on Computer Vision and Pattern Recognition, June 7-12, Boston, USA*, pp. 5353–5360, 2015.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of CVPR 2016 – Conference on Computer Vision and Pattern Recognition, June 26-July 1, Las Vegas, USA*, pp. 770–778, 2016.

[26] T. Tan, Y. Qian, H. Hu, Y. Zhou, W. Ding, and K. Yu, "Adaptive very deep convolutional residual network for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1393–1405, 2018.

[27] X. Shi, M. Zhu, and X. Du, "End-to-end residual CNN with L-GM loss speaker verification system," in *Proceedings of DSP 2018 – 23rd IEEE International Conference on Digital Signal Processing, November 19-21, Shanghai, China*, 2018.

[28] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proceedings of ICLR 2016 – 4th International Conference on Learning Representations, May 2-4, San Juan, Puerto Rico*, 2016.

[29] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," *Neurocomputing*, pp. 227–236, 1990.

[30] Google Developers, "Machine Learning Crash Course - Classification: Accuracy." https://developers.google.com/machine-learning/crash-course/classification/accuracy.

[31] X. D. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall, 2001.

[32] A. G. Katsiamis, E. M. Drakakis, and R. F. Lyon, "Practical gammatone-like filters for auditory processing," *EURASIP Journal on Audio Speech and Music Processing*, vol. 4, pp. 1–16, 2007.

[33] J. Gibson, M. V. Segbroeck, and S. Narayanan, "Comparing time-frequency representations for directional derivative features," in *Proceedings of INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association, September 14-18, Singapore*, pp. 612–615, 2014.

[34] M. Huzaifah, "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks," *CoRR*, vol. abs/1706.07156, 2017.

[35] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America*, vol. 8, pp. 185–190, 1937.

[36] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *Proceedings of SMC 2010 – 7th Sound and Music Computing Conference, July 21-24, Barcelona, Spain*, 2010.

[37] M. Vetterli and C. Herley, "Wavelets and filter banks: theory and design," *IEEE Transactions on Signal Processing*, vol. 40, pp. 2207–2232, 1992.

[38] J. Youngberg and S. Boll, "Constant-Q signal analysis and synthesis," in *Proceedings of ICASSP 1978 – 3rd IEEE International Conference on Acoustics, Speech and Signal Processing, April 10-12, Tulsa, USA*, 1978.

[39] B. C. J. Moore, *Hearing (Handbook of Perception and Cognition, Second Edition)*. Academic Press, 1995.

[40] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," pp. 18–24, 01 2015.

[41] A. H. Moore, J. M. de Haan, M. S. Pedersen, P. A. Naylor, M. Brookes, and J. Jensen, "Personalized signal-independent beamforming for binaural hearing aids," *Journal of the Acoustical Society of America*, vol. 145, pp. 2971–2981, 2019.

[42] F. Chollet *et al.*, "Keras." https://keras.io, 2015.

[43] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[44] N. Gershenfeld, "An experimentalist's introduction to the observation of dynamical systems," in *Directions in Chaos — Volume 2*, pp. 310–353, World Scientific, 1988.

[45] TensorFlow.org Tutorials, "Simple audio recognition." https://www.tensorflow.org/tutorials/sequences/audio_recognition.

[46] N. Blachman and R. Machol, "Confidence intervals based on one or more observations," *IEEE Transactions on Information Theory*, vol. 33, pp. 373–382, 1987.

[47] S. Fernández, A. Graves, and J. Schmidhuber, "An application of recurrent neural networks to discriminative keyword spotting," in *Proceedings of ICANN 2007 – 17th International Conference on Artificial Neural Networks, September 9-13, Porto, Portugal*, pp. 220–229, 2007.