mathematics

MDPI

Article

Acoustic Gender and Age Classification as an Aid to Human–Computer Interaction in a Smart Home Environment

Damjan Vlaj and Andrej Zgank *

Faculty of Electrical Engineering and Computer Science, University of Maribor, 2000 Maribor, Slovenia
* Correspondence: andrej.zgank@um.si; Tel.: +386-2-220-7206

Abstract: The advanced smart home environment presents an important trend for the future of human wellbeing. One of the prerequisites for applying its rich functionality is the ability to differentiate between various user categories, such as gender, age, speakers, etc. We propose a model for an efficient acoustic gender and age classification system for human–computer interaction in a smart home. The objective was to improve acoustic classification without using high-complexity feature extraction. This was realized with pitch as an additional feature, combined with additional acoustic modeling approaches. In the first step, the classification is based on Gaussian mixture models. In the second step, two new procedures are introduced for gender and age classification. The first is based on the count of the frames with the speaker's pitch values, and the second is based on the sum of the frames with pitch values belonging to a certain speaker. Since both procedures are based on pitch values, we have proposed a new, effective algorithm for pitch value calculation. In order to improve gender and age classification, we also incorporated speech segmentation with the proposed voice activity detection algorithm. We also propose a procedure that enables the quick adaptation of the classification algorithm to frequent smart home users. The proposed classification model with pitch values has improved the results in comparison with the baseline system.

Keywords: acoustic classification; acoustic signal processing; Gaussian mixture model; pitch analysis; smart home

MSC: 68T10

Citation: Vlaj, D.; Zgank, A. Acoustic Gender and Age Classification as an Aid to Human–Computer Interaction in a Smart Home Environment. *Mathematics* **2023**, *11*, 169. https://doi.org/10.3390/math11010169

Academic Editor: Daniel-Ioan Curiac

Received: 24 November 2022
Revised: 21 December 2022
Accepted: 26 December 2022
Published: 29 December 2022

1. Introduction

The intensive development of information communications technology (ICT) has spread into all sections of everyday life, including the human living environment. Real-life smart home systems already include successful automation and control support for the variety of scenarios that human users are confronted with. Currently, the majority of smart home users belong to the category of early adopters, but it is expected that the future development of the technology will increase its broad acceptance in the general population [1,2].

An important functionality in a smart home environment is the detection of the user's presence in a room. This can be fulfilled in different ways, focusing on non-invasive methods, wherein users do not need to wear any dedicated device. One of the traditional methods is passive infrared (PIR) motion detection, which yields relatively simple and robust sensors. The disadvantage of this technology is its inability to distinguish between different user categories, such as gender, age, speakers' identity [3], etc. Another method that cannot cope with user categories is speech activity detection (SAD) [4], which provides the smart home system with information about a user's presence solely from the captured speech signal.

Human–computer interaction (HCI) can, in advanced smart home environments, provide rich functionality if it can differentiate between various user categories. To distinguish

between them, machine learning classification can be used. The classification accuracy and, in particular, system complexity largely depends on the category characteristics and intra-category variance. The decision regarding which type of user classification (e.g., gender) to apply is based on HCI scenarios and requested functionalities that need to be applied in the smart home environment. The resulting HCI system in a smart home environment can act accordingly, applying user scenarios, adapting functionality, or deploying entity personalization. The area of personalization [5] covers a large number of possible scenarios. Some smart home entities that are frequently included in the personalization process include the user interface, media and users' content, a recommendation system, AAL functions, etc. The main personalization objective is to achieve better usage acceptability and higher quality of experience. To be able to carry out such category-based presence detection, a more sophisticated approach must be used than IR motion detection or SAD. One possibility is to use image processing [6], and another is to use more complex audio processing. The advantage of presence detection using audio as a modality is its lower computational complexity, lower cost, and better acceptability among users. The user's acceptability is tightly connected with the data privacy question. In the case of audio, a well-considered design can lead to local processing, without the need to use cloud-based speech processing services.

Based on the above characteristics of the advanced HCI interface, we propose an acoustic classification model that determines the age and gender of the speaker from the captured speech signal. Such a classification approach can be used in a smart home environment for precise presence detection. Broadly accepted smart home usage scenarios were analyzed and a decision was made to classify users into three categories: male, female, and child. This represents an effective combination of all speakers' characteristics. A special characteristic of the proposed model is the inclusion of the category of children in the classification, which is not typically represented in speech technologies. This aspect is important due to the smart home user interface design and content-processing process. In the case of children, the personalization steps must be more intensive and age-oriented, emphasizing domain control and the adequacy of the information available to them. The first objective of the proposed acoustic classification model is to provide accurate performance for all three defined categories using pitch processing. Pitch is one of the signal processing values that can contribute most extensively to classification accuracy, as it depends significantly on the speakers' characteristics. We propose an algorithm for efficient pitch calculation. We also propose two-step solutions for including pitch in the classification to enable adaptation in the second step. The second objective of the proposed model is to simplify the development of acoustic age and gender classification for presence detection in a smart home environment. The motivation was to reuse available speech recognition modules from a smart home environment. This results in less complex speech technology methods that can even be used by resource-constrained devices.

Using speech as an input modality has both advantages and disadvantages. The speech signal propagates through the room, which means that the capture devices need not focus directly on the user. This can significantly improve the system's usability. There are two shortcomings present for speech modality. The first one is the sensitivity of the speech signal in relation to other audio sources co-existing in the room. The disturbing audio sources' types vary according to the scenario: background music/TV, home appliances, other speakers, domestic animals, street noise, etc. The second one is the issue of the privacy of uttered information, which can be successfully handled by local processing in the scope of the embedded systems. The speech-based presence detection system can be combined with other presence detection entities, such as PIR motion detection sensors. This can result in improved performance, as the modalities of signals and noises differ. The end result is the limiting of shortcomings connected with the pure speech-oriented system.

The paper is organized as follows. Section 2 presents a literature review from two perspectives: gender classification and smart home systems. Section 3 first presents the proposed model and then describes the theory of acoustic gender classification, with an

emphasis on approaches used in the experiments. Then, the system design applied in the experiments is presented. The speech processing results are presented in Section 4. The discussion is provided in Section 5, and the conclusion is provided in Section 6.

2. Literature Review

The objective of this work was to establish a back-end smart home service, which could be used for detecting the presence of users solely via speech modality and classify them accordingly. Thus, the addressed related work also covers two fields. The first one is the field of digital signal processing and gender classification from speech. The second one is the field of smart home systems and services.

The topic of acoustic gender classification is an area in spoken language technology with a long history. The first systems emerged decades ago [7–9], mainly as sub-components of automatic speech recognition systems. The basic idea—of how to detect gender from acoustic signals—is usually pursued via the spectral and temporal characteristics of the captured speech signal [10]. To be able to carry out gender classification, two approaches need to be combined: first, the representative features are extracted from the captured audio signal [11], and second, the appropriate machine learning approach is used to classify them [12].

The acoustic feature extraction procedure is a key factor in successful gender classification. Various approaches, such as mel-frequency cepstral coefficients (MFCC) [13,14], pitch [15], and RASTA [15], have been used for the gender classification task. In general, MFCC feature extraction usually provides good classification results. The speech rate, pauses, loudness, intonation, and voice quality can be categorized as paralinguistic features. Similar to acoustic features, paralinguistic features can also be used for gender classification [16], with the objective of broadening the data available for classification. This can improve the overall classification accuracy.

An important issue in the case of gender classification from speech is its robustness against the acoustic background and other degradation events. The background signal and noise can reach high energy levels and, consequently, significantly disturb system operation in smart home scenarios. Islam [17] showed that GFCC features also significantly improve the robustness and effectiveness of gender classification in a harsh environment.

One of the baseline approaches to addressing gender classification is Gaussian mixture models (GMMs) [18]. The GMM gender classification approach can show high precision with relatively low complexity, which is important for smart home scenarios, where limited embedded resources are frequently available. Ranjan et al. [19] also showed that GMM gender classification achieves good results in different languages, or even in a multilingual environment.

Hidden Markov models (HMMs) have been used for gender and age classification [20], and also for the classification of various human activities in natural environments [21]. The use of HMMs introduces another model's architecture to the classification task, which can improve the robustness, accuracy, and reusability of real-life systems. The combination of several statistical approaches is presented in [22], where universal modeling (UM) based on GMM clustering was used.

Another machine learning approach used for gender classification is support vector machines (SVMs). Bocklet et al. [23] showed that SVM can achieve high-accuracy gender recognition results. The i-vector approach proposed by Dehak [24] was also applied successfully for gender classification in complex spoken scenarios [19].

Deep neural networks (DNNs) were used for gender classification by various authors [25,26]. The main objective was to improve accuracy and combine the gender classification system with the main automatic speech recognition system using the same architecture. Prior to DNNs, other neural network methods, such as multi-layer perceptrons (MLPs) [27], were also successfully implemented for gender classification.

The majority of gender classification systems found in the literature only deal with adult speech, thus classifying between males and females (and unknown). In the case

of a smart home, distinguishing between adult and child can also be important. The Paralinguistic Special Session of Interspeech 2010 [28] addressed this topic. The aGender speech database [29], which is applied for the classification task, originated from long conversational telephone sessions, and the speakers were classified into three gender categories and seven combined categories. Meinedo and Trancoso [30] presented a system that used a combination of four different corpora with the fusion of acoustic and prosodic features, and this was able to classify gender with an 84.3% average recall. Yücesoy and Nabiyev [29] carried out gender classification on the aGender speech database with a combination of three subsystems at the score level. The experimental system provided a 90.39% classification success rate for the gender category. This result shows that there are still challenges in the area of gender classification when children's speech is incorporated into experiments.

With the development of algorithms, systems, and terminal equipment, the number of possible use cases increased, and nowadays, gender classification systems can be used successfully in the smart home environment [31]. Speech activity detection, which can be seen as a simplified gender classification approach for the smart home, was addressed by SASLODOM, part of the EVALITA 2014 challenge [32], wherein three different SAD systems were presented. The best system achieved a 2.0% SAD error rate at the frame level, which is already usable in real-life scenarios applicable to SAD. Gender classification can also be helpful for social robots as part of the smart home environment [33]. The availability of extended speech databases also enables a combined approach, whereby age and gender were processed in parallel [34,35].

The literature review presents a general overview of approaches to carrying out the gender classification task. In our work, the emphasis will be placed on particular solutions for acoustic presence detection, as well as gender and age classification, in the smart home environment, where the combination of accuracy and required system resources plays an important role.

3. Materials and Methods

This section presents the proposed procedure used for gender and age classification from input speech signals. First, we present the entire process of preprocessing and extracting the necessary information from the input speech signal so that, in the end, we can determine the presence of a male, female, or child in the environment through classification procedures. Then, we present, in more detail, the voice activity detection (VAD) algorithm and the procedure for determining the pitch value from the input speech signal. The pitch value of an individual speaker gives essential information about whether the speaker is a male, female, or child. Therefore, we chose the pitch value as one of the more essential features in our proposed procedure. We presented the VAD algorithm and the pitch value determination in one of our previous works [36]. We enhance the procedures in this paper to improve the algorithm's performance, which we also describe in more detail in this section's second and third subsections. Next, we present the feature extraction algorithm included in our setup. The training of Gaussian mixture models (GMMs) is presented thereafter.

3.1. Proposed Gender and Age Classification

Here, we present the proposed gender and age classification procedure in detail. Figure 1 will form the basis for describing the details of the proposed procedure. We want to extract information about a person's gender (male or female) and age (adult or child) from the speech signal spoken by a person present in an intelligent environment. The input speech signal is divided into overlapping frames. All further information extraction is from the frames. In the speech signal, most information about gender and age is present in the voiced frames of the speech signal. The voice activity detection (VAD) algorithm detects the voiced frames in the speech signal. The voiced frames of the speech signal are the basis for determining the speaker's pitch value. The next step in the procedure is calculating

the 12 mel-frequency cepstral coefficient (C1–C12) features and energy, as specified in the standard [37]. After that, we carried out the composition of the feature vectors that were finally used. Because we did not want to change the size of the feature vector, we decided to replace coefficient C12 with the pitch value. To improve the effectiveness of gender and age classification, we have also calculated the first and second derivatives of the feature vector coefficients. Once we derived the final feature vector and the pitch value for each frame, we began classifying the person's gender and age.

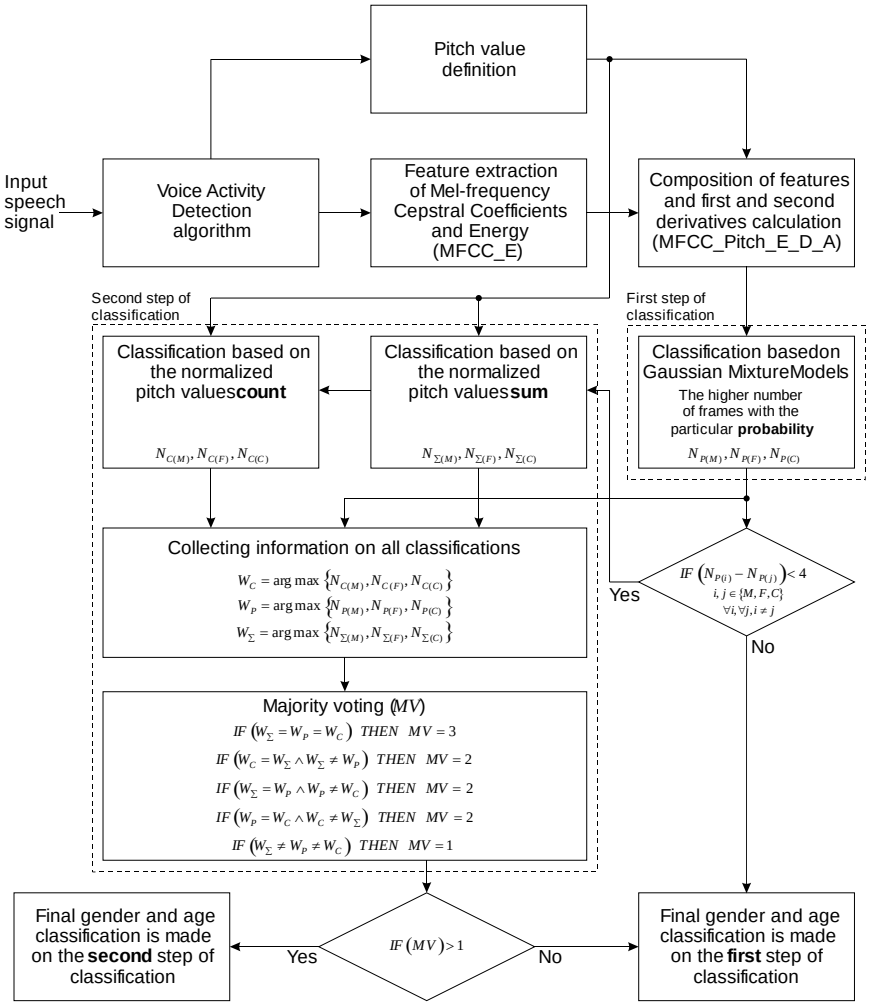

Figure 1. The proposed gender and age classification procedure.

We propose a gender and age classification procedure comprising two steps [38]. We decided on a two-step classification process because, in the second step, based on the analysis of the pitch value of the speakers present in an intelligent environment, we can improve the classification. In the first step, gender and age classifications are based on Gaussian mixture models (GMMs). For each frame, three probabilities, $P(M)$, $P(F)$, and $P(C)$, are determined, representing the probabilities of a male, a female, or a child. The higher probability determines to which gender or age the frame belongs. Next, the frame counters ($N_{P(M)}$, $N_{P(F)}$, and $N_{P(C)}$) are determined for all three classification categories. For each frame, only one frame counter is incremented; that is, the one whose frame has a higher probability. The frame counter with the higher value indicates which gender or age the entire input speech signal belongs to, and it is presented as the winner, W_P, of the classification based on GMMs. We completed the classification process if any frame counter value stood out and was at least four counts greater than the other two. In this case, we considered the GMM-based classification as a final conclusion of gender and age

classification. However, if the difference between the two frame counters was less than four, we continued with the second classification step. The right decision block in Figure 1 presents this decision.

In the second step, we propose two procedures of gender and age classification. These are (a) classification based on the normalized pitch value counts and (b) classification based on the normalized pitch value sums. The normalized pitch values are obtained in both procedures by analyzing the speech recordings used to train the GMMs. All frames in which we can determine the pitch value were used for analysis. First, we established three groups of male, female, and child speakers. After that, we divided the pitch values into intervals of 10 Hz. In the 80 Hz interval, there are pitch values between 80 and 90 Hz; the 90 Hz interval covers all pitch values between 90 and 100 Hz, etc. The male speakers' recordings exhibited the most pitch values between 110 and 120 Hz. All the counted pitch values from the male, female, and child speakers were normalized according to the highest value observed by the male speaker.

For the classification based on the normalized pitch value counts, we analyzed the occurrence of the normalized pitch values for the male, female, and child speakers. For each frame, three count areas, $C(M)$, $C(F)$, and $C(C)$, were determined, representing the areas for a male, a female, or a child speaker. If the normalized pitch value for a current frame was below 160 Hz, it belonged to the count area $C(M)$; if it was between 160 and 230 Hz, it belonged to the count area $C(F)$, and if it was above 230 Hz, it belonged to count area $C(C)$. Here, we also determined the frame counters ($N_{C(M)}$, $N_{C(F)}$, and $N_{C(C)}$) for all three classification categories. For each frame, only one frame counter was incremented—the one whose frame belongs to a particular count area. The frame counter with the higher value was used to determine to which gender or age group the entire input speech signal belonged and was presented as the winner, W_C, of the classification based on the normalized pitch value counts.

The classification based on the normalized pitch value sums also uses the analysis results. We have defined three sums: for male $N_{\Sigma(M)}$, female $N_{\Sigma(F)}$, and child $N_{\Sigma(C)}$ speakers. For each frame in which we could detect the pitch value in the speech signal, we added the normalized values obtained from the analysis to the sums of each speaker. For example, a pitch value of 173 Hz was determined within a particular frame. This pitch value was between 170 and 180 Hz, so a 170 Hz interval was selected for all classification groups. Consequently, the normalized value of 0.14 was added to the $N_{\Sigma(M)}$, the normalized value of 0.37 was added to the $N_{\Sigma(F)}$, and the normalized value of 0.08 was added to the $N_{\Sigma(C)}$. Here, can be seen that a normalized value of 0.37, which was added to the $N_{\Sigma(F)}$, was the largest compared to the other two values. This is understandable since this normalized value is located between 160 Hz and 230 Hz, which belongs to the female speakers' count area $C(F)$. The most significant sum value of three sums ($N_{\Sigma(M)}$, $N_{\Sigma(F)}$, and $N_{\Sigma(C)}$) was used to classify to which gender or age the entire input speech signal belonged, and was presented as the winner, W_Σ, of the classification based on the normalized pitch value sums.

The second classification step ends with collecting information about the winners (W_C, W_P, and W_Σ) of all three described classification procedures and majority voting, MV. We performed all three classification procedures on the same speech signal. The results of all three might be the same, but sometimes they give different results. When the results are the same, the majority vote equals three. Then, all classification procedures can be used to determine whether the speaker in the recording is a male, a female, or a child. In such a case, the final decision of the second classification step is simple. If the majority vote is equal to two, this means that at least two processes give an identical classification. In this case, the final classification is the same as the majority vote winner. However, if all three classification procedures give different results, the majority vote is equal to one. In such a case, the final classification of gender and age is based on the first classification step or GMM-based classification.

We will use the proposed gender and age classification system in a smart home environment. There is always the question of how to update and improve such a system.

In this paper, we propose another procedure that would allow for the fast adaptation of the system to users who appear frequently in a smart home environment. The idea of the procedure is based on the fact that the classification based on GMM remains the same. It means that the test set does not influence the trained GMMs. The change is that the system monitors the correctness of the classification and records pitch values on the basis of the frame of the user in the smart home environment. When we have a sufficiently large number of captured pitch values for users, we use these values to adapt the gender and age classification system to them. This sufficiently large number of pitch values can be taken from the analysis. An experiment has been performed to confirm the adaptation procedure, and the results are provided in Section 4.

3.2. Voice Activity Detection Algorithm

An essential contribution of the voice activity detection (VAD) algorithm is real-time noise energy estimation. Such an algorithm can be used in smart home environments where the noise level can vary significantly. Frame energy and zero-crossing measures are used for the VAD on each acoustic frame. A speech signal is cut into 50 percentage overlapped frames with durations of 25 milliseconds. Owing to the fact that the speech signal is sampled with a frequency of 16 kHz, the duration of each frame can be presented as a 400-sample window. To explain the process of VAD decision determination, we will use Figure 2. The frequency spectrum (Figure 2a) and signal representation in the time domain (Figure 2b–f) of the specific values are provided for a captured spoken sample in which the digit sequence "seven six one three" is uttered.

The frame energy, E_f, values are presented as a blue line in Figure 2b. The frame energy, E_f, value is calculated as in (1) from the N samples of the input signal, s.

$$E_f = \frac{\sum_{i=1}^{N} (s[i])^2}{N} \tag{1}$$

We did not use the logarithmic function in calculating the frame energy, because it would be more difficult to define the threshold that determines the presence of speech. The noise area energy values, E_n, are presented in Figure 2c. The value E_n is calculated for each frame as in (2) from the 10 values of the cyclic noise buffer, N_{buff}.

$$E_n = \frac{\sum_{i=0}^{9} N_{buff}[i]}{10} \tag{2}$$

The value E_n is calculated as an average value within a cyclic noise buffer, N_{buff}, that contains the frame energy, E_f, of the last 10 noisy frames.

The energy values, E_f, of the first 10 frames in the captured audio signal are mapped into the cyclic noise buffer, N_{buff}. After the first 10 frames, only the noisy frames with the weighted energy values, E_f, are added in the last place of the cyclic noise buffer, N_{buff}. The decision regarding which energy value, E_f, contains noise is presented in

$$N_{buff}[9] = \begin{cases} E_f; & E_f \leq 2 \cdot E_n \\ E_f/2; & 2 \cdot E_n < E_f \leq 4 \cdot E_n \\ E_f/4; & 4 \cdot E_n < E_f \leq 8 \cdot E_n \end{cases} \tag{3}$$

As can be seen in (3), the cyclic noise buffer, N_{buff}, is not updated when the current frame energy value, E_f, is 8 times larger than the noise area energy value, E_n, in the same frame. This limit has been determined empirically. If the cyclic noise buffer, N_{buff}, is not updated, then the noise area energy value, E_n, is also not updated according to (2). This is also presented in Figure 2c. This coincides with the beginning of the speech occurrence (digit "seven") in the captured audio signal and can be seen in Figure 2a.

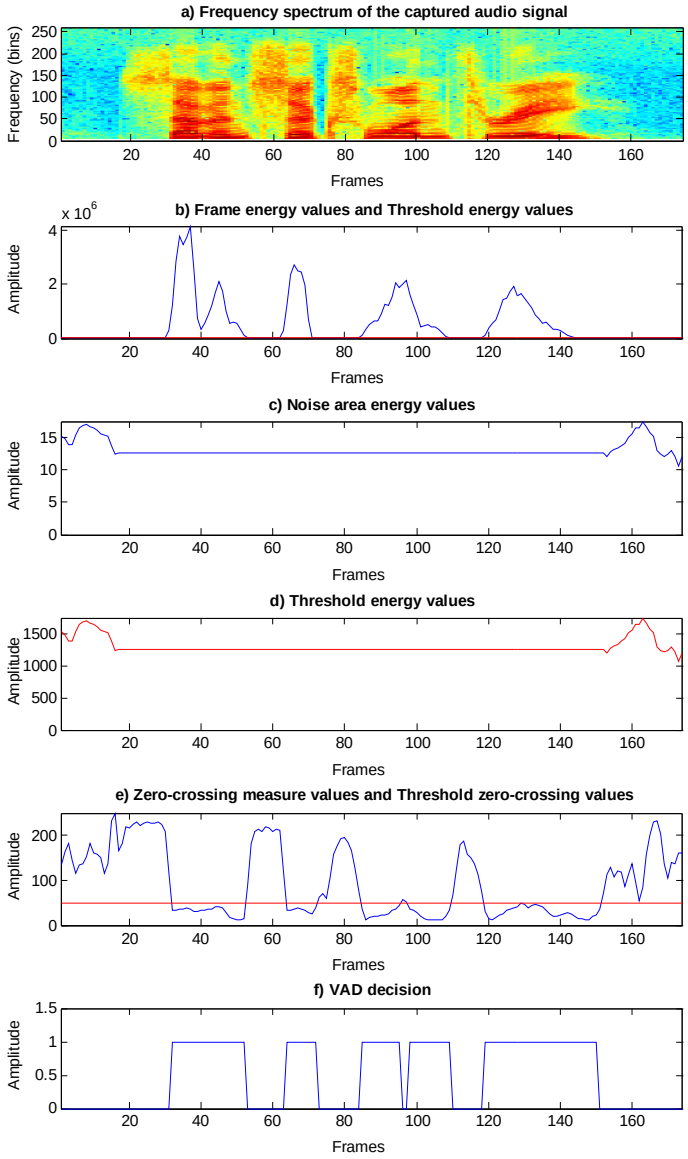

Figure 2. VAD decision determination: (**a**) Captured audio signal frequency spectrum. (**b**) Frame energy values and threshold energy values. (**c**) Noise area energy values. (**d**) Threshold energy values. (**e**) Zero-crossing measure values and threshold zero-crossing values. (**f**) VAD decision.

The next step in determining the VAD decision is to determine the threshold energy value, E_{Th}. The threshold energy value, E_{Th} is calculated as in (4) with the help of factor, f, and the noise area energy value, E_n.

$$f = \begin{cases} 100; & E_n \leq 100 \\ 100 - 0.1 \cdot E_n; & (E_n > 100) \ \wedge \ (E_n < 900) \\ 10; & E_n \geq 900 \end{cases} \tag{4}$$

$$E_{Th} = f \cdot E_n$$

The achieved result can be seen in Figure 2d. It is evident from the decision procedure that we used a different factor value, f, to determine the threshold energy value, E_{Th}. If the noise area energy value, E_n, is small (smaller than or equal to 100), it is necessary to raise the threshold energy value, E_{Th}. If the value of factor f is 10, then a slight increase in frame energy value E_f would lead to the wrong VAD decision, since the zero-crossing values

(Figure 2e) in the non-speech areas are also large. Therefore, we need to use a larger factor value, f, (in our case it is 100) so that wrong VAD decisions are less probable. On the other hand, if the noise area energy value, E_n, is large (larger than or equal to 900), it is necessary to reduce the threshold energy value, E_{Th}. Such high energy values of E_n occur if there is no silence at the beginning of the captured audio signal and speech occurs immediately. If the value of factor f is 100, then the threshold energy value, E_{Th}, would be too high, which would mean that the VAD algorithm would not detect the voiced speech segments in the captured audio signal. Therefore, in this case, we set a smaller factor value, f, (in our case, 10) so that the VAD algorithm can detect the speech segments in the captured audio signal. The high noise area energy value, E_n, decreases as soon as the conditions in (3) are met. However, if the noise area energy value, E_n, is between 100 and 900, then the value of factor f, as well as the threshold energy value, E_{Th}, changes linearly according to (4). Typically, for the captured audio signal, the time domain representations of the noise area energy values, E_n, (Figure 2c) and the threshold energy values, E_{Th}, (Figure 2d) are identical but multiplied by the constant factor, f, used. The time domain representations of the threshold energy values, E_{Th}, in Figure 2d are presented with a red line. The same value is also presented with a red line in Figure 2b.

We can derive additional information for better VAD decisions from the zero-crossing measure value, ZC_m. The enormous zero-crossing measure value in the frame represents the frame containing noise, or unvoiced speech, in the audio signal. For example, consonants in the speech signal belong to unvoiced speech. Figure 2a shows the frequency spectrum of the digit sequence "seven six one three", and the words seven and six contain the consonant "s". In word seven, the consonant "s" is present from frame 18 to frame 30, while in word six it is present from frame 53 to frame 63, and from frame 75 to frame 84. The value of the zero-crossing measure is presented as a blue line in Figure 2e. The ZC_m values in the unvoiced speech and noise signal regions are large and much more significant than those in a voiced speech signal region. The zero-crossing threshold value, ZC_{Th}, determines the segments of unvoiced speech and segments of the voiced speech signal. We set this value to 50, which is presented in Figure 2e as a red line. As mentioned before, one frame contains 400 samples. Having the value ZC_{Th} set at 50 means that the signal crosses the zero value at every 8 samples. This also means that the signal reaches its positive peak at every 16 samples. In this case, for a sampling frequency of 16 kHz, the pitch value would be 1000 Hz, which is almost impossible.

The proposed VAD decision is calculated from frame energy value E_f, zero-crossing measure value ZC_m, threshold energy value E_{Th}, and zero-crossing threshold value ZC_{Th}, as presented in (5). For each frame, the VAD decides if it contains voiced speech or not. Figure 2f shows the VAD decision on the captured audio signal. Voiced frames are then used for pitch value determination.

$$VAD = \begin{cases} 1; & (E_f > E_{Th}) \wedge (ZC_m < ZC_{Th}) \\ 0; & (E_f > E_{Th}) \wedge (ZC_m \geq ZC_{Th}) \\ 0; & E_f \leq E_{Th} \end{cases} \quad (5)$$

3.3. Pitch Definition

A pitch value, or speaker's fundamental frequency, can be determined from the speech signal's time domain representation or the frequency spectrum. Our pitch determination process is based on a periodic pattern, which can be found in the time domain representation of the speech signal. A repeating periodic pattern can be found in all vowels, sonorant consonants (/n/, /m/, /l/, etc.), and also in voiced obstruents (such as /b/, /d/, /g/) [39]. To facilitate the interpretation of the pitch determination process, we will use time domain representation of the vowel /i/ in word six of the captured audio signal with the digit sequence "seven six one three". The frequency spectrum of this digit sequence is presented in Figure 2a. The time domain representation of the 65^{th} frame of this sequence is presented in Figure 3a. The blue line represents speech signal samples, and we can see a repeating

periodic pattern. In the next paragraph, we will present the procedure by which the peaks are detected in each frame. The pitch can then be calculated from the difference between correct peaks.

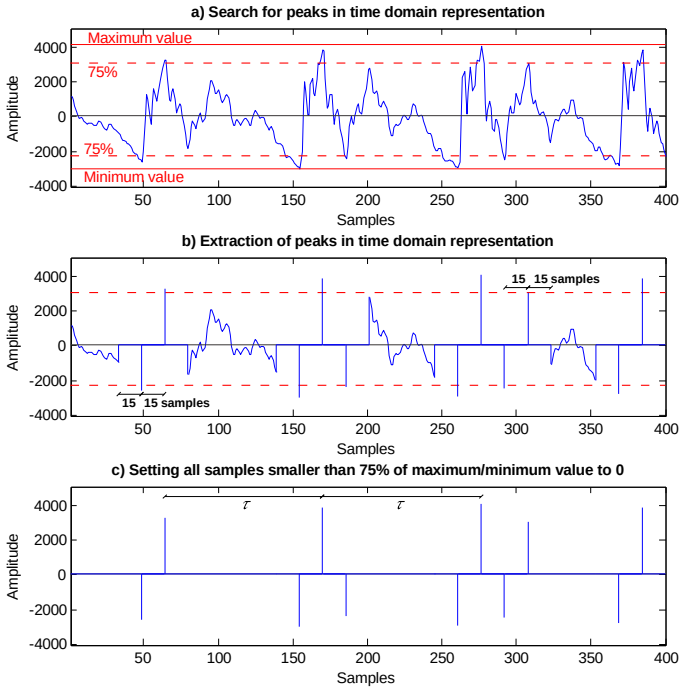

Figure 3. The time domain representation of the 65th frame in the digit sequence "seven six one three": (**a**) Search for peaks in one speech signal frame. (**b**) Extraction of peaks, where 15 samples left and right of the peak are set to 0. (**c**) All samples smaller than 75% of maximum/minimum value are set to 0.

When we define pitch value, we must first define the highest maximum value between positive samples' values and the lowest minimum value between negative samples' values. The samples' highest maximum and lowest minimum values in the frame are presented as red lines in Figure 3a. After that, we must define positive and negative peaks. The current peak maximums or minimums are detected in samples where greater than 75% of the maximum or minimum value is detected in the frame. The maximums are searched from the highest maximum to 75% of their value. The 15 samples left and right of the positive or negative peaks are set to 0. Figure 3b shows the result of this procedure. In the end, all other samples below 75% of the highest maximum or lowest minimum are set to 0. Figure 3c shows this result. If we look at the positive peaks that we have found, we can see that the first, second, third, and fifth are detected correctly. The fourth positive peak is incorrect. For negative peaks, two peaks (third and fifth) are defined incorrectly.

Finding the difference or the number of samples between the peaks is the next step in the procedure. As can be seen in Figure 3c, the difference is represented by the variable τ. Differences are calculated between all adjacent peaks. Positive and negative peaks' positions and the calculated differences between adjacent peaks can be seen in Table 1. Only the differences between the first and the second peak and between the second and the third peak of the positive peaks, and the difference between the first and the second peak of the negative peaks, gave correct results that could be used to determine the correct pitch value. After the calculation, the differences are sorted from the highest to the lowest value. To determine the pitch, the maximum calculated difference is used, along with those that deviate from this value by 10% or less. In the presented frame, only the first two differences from the positive peaks are used (see Table 1, values 106 and 107), as well as the first difference from the negative peaks (see Table 1, value 106). The average difference is determined from the differences that are used. The fundamental frequency, F_0, or pitch

value can be calculated as in (6), where f_{samp} is the sampling frequency and $\overline{\tau}$ is the average difference between the peaks.

$$F_0 = \frac{f_{samp}}{\overline{\tau}} \tag{6}$$

Table 1. Positive and negative peaks' positions in the 65th frame in the digit sequence "seven six one three" and differences calculated between adjacent peaks.

Positive peak position	**Difference between adjacent peaks**
64	106
170	107
277	31
308	76
384	
Negative peak position	**Difference between adjacent peaks**
49	106
155	31
186	75
261	31
292	77
369	

This method of determining a pitch is only applicable when a repeating periodic pattern is detected in the speech signal. In any case, such a signal is not present when the VAD algorithm does not detect the presence of a voiced segment in the speech signal. When we know the pitch of the speech signal's voiced segments, then this information can be used to detect the speaker's gender in two ways. One is that we use different levels of thresholds for determining a speaker's gender, while the other is to use the pitch as a feature for training different models. In this paper, Gaussian mixture models (GMMs) are used for the model training process. When training GMMs, specific cases may arise wherein it is not possible to train them if all the training values are not defined. In the noisy or silent segments in the audio signal, and in the unvoiced speech segment, we do not have information about the pitch, because it cannot be determined in these segments. Now the question concerns which value to set in these audio signal segments. Our approach here is that these values should be smaller than the value of the pitch that may occur in the voiced part of the speech signal. We decided that this value should be less than 40 Hz. For the modeling process, it is not appropriate that this value be constant for the whole unvoiced speech signal segment. Therefore, we determined the apparent value of the fundamental frequency, F_0, or pitch value as in (7), where F_{max} is the maximum apparent value of the pitch, *frameLength* is the length of the frame, and *averagePeak* is the average value of the peaks' positions in the frame.

$$F_0 = \frac{F_{max} \cdot averagePeak}{frameLength} \tag{7}$$

From Figure 4, we can determine these values. If the value *frameLength* is 400, the value F_{max} is set to 40, while the value of *averagePeak* is calculated by summing the positions of the three positive peaks (218, 329, and 348) and the positions of the two negative peaks (201 and 313). This gives the calculated average value of 281.8. The apparent pitch value is then taken as 28.18 Hz.

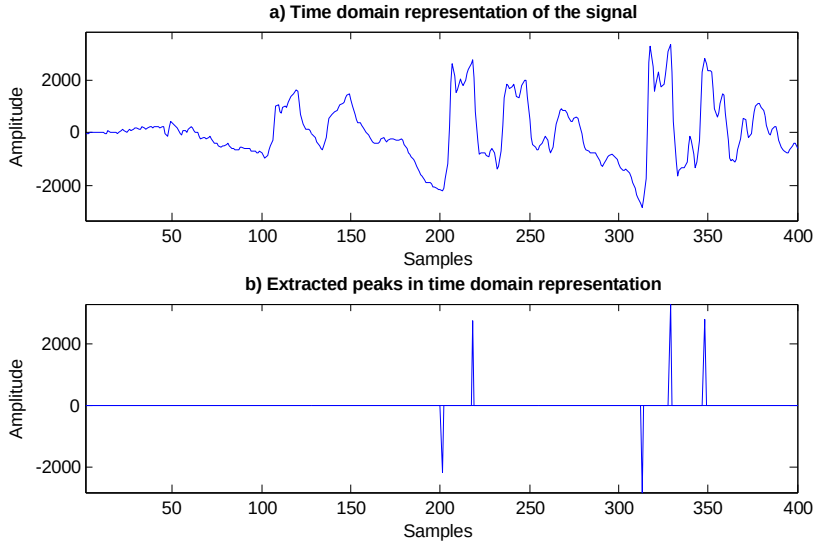

Figure 4. The time domain representation of the frame at the boundaries of the transition from the unvoiced to the voiced segment of the audio signal: (**a**) The time domain representation of the signal. (**b**) Extracted peaks in the time domain representation.

The proposed Equation (7) makes it possible to determine the apparent pitch in the unvoiced speech signal segments. We also used (7) when we could not define the pitch in the voiced speech signal segment. This occurs at the boundaries of the transition from the voiced to the unvoiced segment of the speech signal, and vice versa. Figure 4 shows an example in which we could not determine the pitch from the detected peaks correctly. We decided to use (7) when the procedure did not detect any positive or negative peaks between 0 and 200, or between 200 and 400 samples. This equation and procedure (7) are also used in cases when we detect the following:

(a) Less than two positive or negative peaks;
(b) Twice as many positive or negative peaks between 0 and 200 as between 200 and 400 samples;
(c) Twice as many positive or negative peaks between 200 and 400 as between 0 and 200 samples;
(d) Two peaks where the difference between them is greater than 200 or less than 25 samples.

3.4. GMM Training

The Gaussian mixture models (GMMs) belong to the group of statistical speech recognition methods that apply the weighted sum of the Gaussian probability density functions as components. Each component is defined by the mean vector, mixture weights, and covariance matrix, which is, in the case of speech processing, frequently diagonal. The GMMs are trained in an iterative way with the Baum–Welch algorithm [40], which applies the expectation-maximization (EM) algorithm to determine the maximum likelihood estimation of the unknown models' parameters on a set of training feature vectors.

To train GMMs, we used 12 mel-frequency cepstral coefficient (C1–C12) features. We replaced the coefficient C12 with the pitch value and added the energy coefficient. We thus derived 13 coefficients in the feature vector. The most significant coefficients' values in the feature vector are the values of the energy coefficient, and these are in the range of 20. However, since pitch values can also be up to 500 and over, we decided to divide the pitch values by 10 so that these coefficient values would not be too high. To improve the effectiveness of gender and age classifications, we have also calculated the first and second derivatives of the feature vector coefficients.

The next step is to determine the number of models we have trained. The VAD algorithm gives us the information wherein the audio recordings are speech signals and

there is also silence. Based on this information, we trained four GMMs. On the speech signal parts, we trained the models for a male, a female, and a child (a boy or a girl), and for the rest of the signal, we trained silence. We used the hidden Markov model toolkit [40] for the GMMs' training. The GMMs' training procedure is provided in Figure 5.

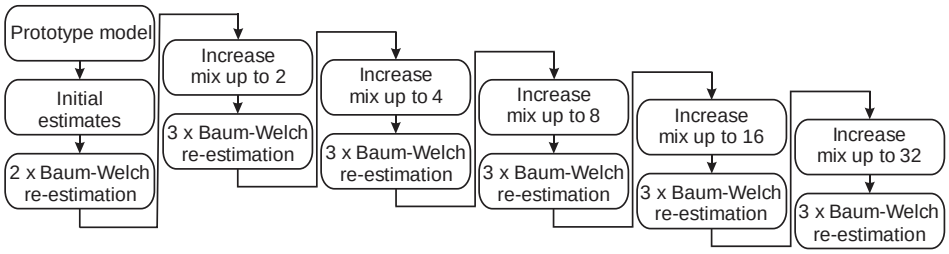

Figure 5. The GMM training procedure with up to 32 Gaussian mixtures per state.

When training GMMs, we started with a prototype model, which defines the required model topology. The topology of a single GMM is presented as a single-state model, and has the form of the required model, except that means are set to 0, variances are set to 1, and mixture weights are set to 1. The next step is to provide initial estimates for the feature vector single model parameters using a set of observation sequences for each model (male, female, child, and silence, if used). The next step is to perform two basic Baum–Welch re-estimations of the single model parameters using a set of observation sequences. Then, we used the procedure to increase the number of Gaussian mixtures. In the following, we again perform Baum–Welch re-estimation of the parameters, but it is now completed three times. The last two steps are repeated all the way to training GMMs with up to 32 Gaussian mixtures.

3.5. Experimental Design

For the experimental design, we used the speaker-independent connected digits American English speech database TIDIGITS [41]. This speech database is widely used for speech technology research, and it is one of the few that also includes utterances spoken by children in an equally balanced way. The original audio recordings were collected in a quiet environment and digitized at a 20 kHz sampling rate. For the needs of this research and the needs of the developed application, which will be used in the smart home environment, we downsampled the original audio recordings to the 16 kHz sampling frequency. The complete speech database contains 326 speakers, of which 111 are male, 114 are female, and 101 are child speakers (50 boys and 51 girls). In our research, we did not separate child speakers by gender. The age of male speakers is between 21 and 70 years, females are between 17 and 59, and children are between 6 and 15 years. The recordings of the speech database are divided into a training and a test set, which is proposed by the authors of the speech database. The training material contains 163 speakers, of which 55 are male, 57 are female, and 51 are child speakers—together representing 12,549 audio recordings. For the test material, 163 speakers remained, of which 56 were men, 57 were female, and 50 were children, which together represented 12,547 audio recordings. The speakers pronounced the following English digits in the recorded database: zero, one, two, ..., nine, and oh. Almost all of the speakers pronounced 77 isolated or connected digits, of which 22 were isolated, and digit sequences two, three, four, five, and seven digits long were uttered 11 times.

4. Results

In this section, we will first present a pitch value analysis of each recording in a speech database using the procedure for determining the pitch value presented in this paper. This analysis helps us in the later interpretation of the results. For those frames in the audio recording for which we were able to determine the pitch value, we compared

their values. The pitch value most often detected is determined as the pitch value of the speaker's voice in the audio recording. The analyses are presented in Figure 6. As can be seen from the analysis, the male speaker's pitch value was the least distributed. Most of the recordings with the male speakers had a pitch value of 110 Hz for both the training and test set materials. The pitch values in recordings with female and child speakers were more dispersed. When we look at the analysis results of the children's recordings in Figure 6c, we can see that most of the speakers in the training set of the audio recordings had slightly higher pitch values than most speakers in the test set of the recordings, where their pitch values were somewhat lower. Thus, the pitch values of the child speakers in the test set recordings were closer to the pitch values of the female speakers that were used in the recordings for the training set. If we look only at the pitch values, we can see frequent confusion between child and female speakers in the gender and age classifications.

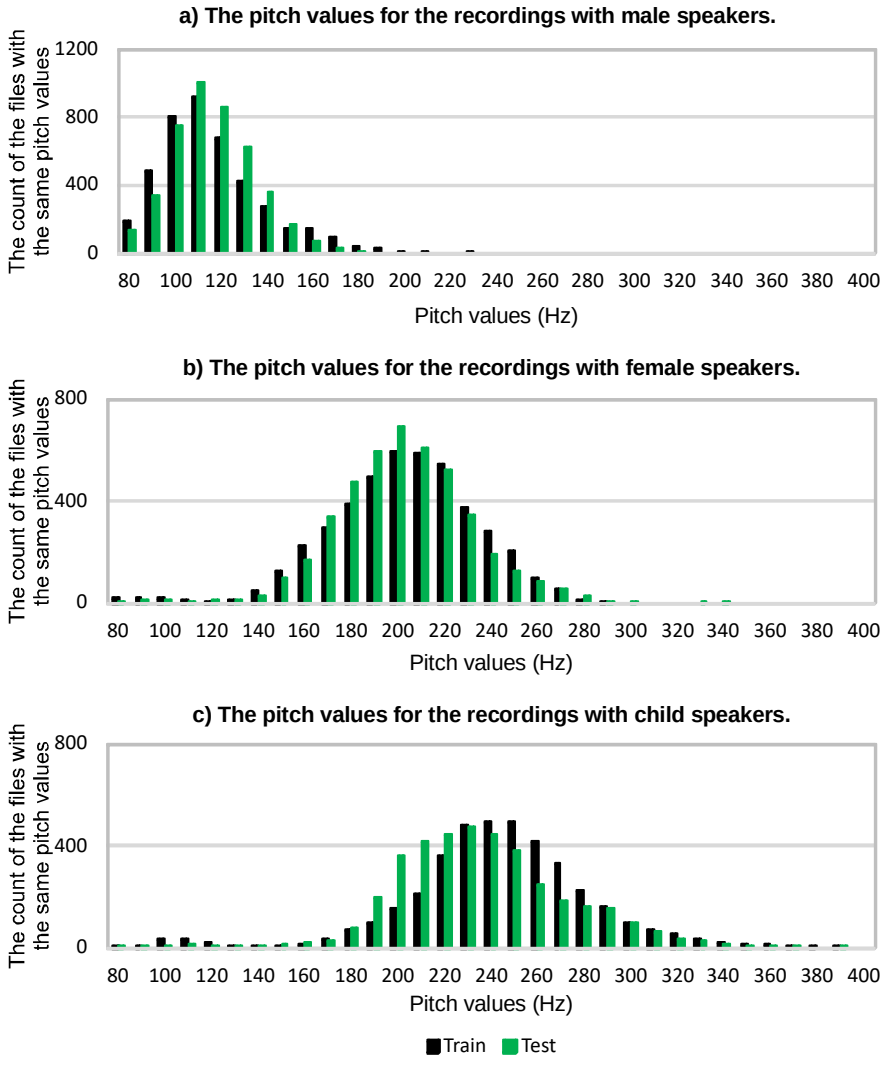

Figure 6. The count of the files with the same pitch values for the recordings with (**a**) male, (**b**) female, and (**c**) child speakers.

We will continue with presenting the results of the speaker's gender and age classification derived from experiments. The accuracy results of gender and age classification, and the statistical analyses of the results with confidence intervals, are presented in Table 2. The associated confusion matrices with recall results are provided in Table 3. The experiments and associated notations used in Tables 2 and 3 are explained in the following list, where (c)–(e) were experiments that were compared with the algorithms presented in the paper:

(a) The classification was based on the count of the pitch values in all frames with calculated pitch values. We performed a classification regarding whether it belongs to a male, female, or child speaker. This decision is presented as a classification based on the normalized pitch value counts. For this classification, we used the notation *Pitch value counts*;

(b) The classification was based on the normalized pitch value counts in all frames in which we could determine the pitch value. We have defined three sums of the normalized pitch value counts for male, female, and child speakers. This decision is presented as a classification based on the normalized pitch value sums. For this classification, we used the notation *Pitch value sums*;

(c) The classification was completed on the basis of the trained three states of the left-to-right monophone HMMs, the topology of which is described elsewhere [40]. In this case, we used three HMMs for male, female, and child speakers and one additional state, a silence HMM. HMMs were trained on the recordings with 11 mel-frequency cepstral coefficients (C1–C11), the pitch value divided by 10, logarithmic energy, and the first and second derivatives of those coefficients. For this classification, we used the notation *Three-state monophone HMMs—MFCC_Pitch_E_D_A*;

(d) The classification was completed on the basis of the trained sixteen states of the word HMMs, the topology of which is presented elsewhere [42]. In this case, we used three sixteen-state HMMs for male, female, and child speakers and one additional 3-state silence HMM. HMMs were trained on the recordings with 11 mel-frequency cepstral coefficients (C1–C11), the pitch value divided by 10, logarithmic energy, and the first and second derivatives of those coefficients. For this classification, we used the notation *Sixteen-state word HMMs—MFCC_Pitch_E_D_A*;

(e) The classification was based on an idea presented in previous work [22], where universal modeling (UM) based on GMM clustering was used. In this case, we used three types of features. The first type of feature was 13 mel-frequency cepstral coefficients (C0–C12); the second type of feature was MPEG-7 low-level descriptors (LLDs), as presented in [21]; and the third type of feature was perceptual wavelet packets (PWP), as presented in [43]. For this classification, we used the notation *Universal modeling based on GMM clustering*;

(f) GMMs were trained on segmented recordings with 12 basic mel-frequency cepstral coefficients (C1–C12), logarithmic energy, and the first and second derivatives of those coefficients. For this classification, we used the notation *GMMs with segmentation—MFCC_E_D_A*;

(g) GMMs were trained on segmented recordings with 11 mel-frequency cepstral coefficients (C1–C11), the pitch value divided by 10, logarithmic energy, and the first and second derivatives of those coefficients. For this classification, we used the notation *GMMs with segmentation—MFCC_Pitch_E_D_A*;

(h) The proposed final gender and age classification was based on a combination of the classifications used in all three experiments (a), (b), and (g). This proposed classification is presented in Section 3.1. For this classification, we used the notation *Proposed algorithm with segmentation*;

(i) The experiment applies in the same way as in experiment (h), with the exception that, in this case, an adaptation of the proposed gender and age classification algorithm is made. With a sufficiently large number of pitch values of the users, we can adapt the normalized pitch value counts and sum. In the last paragraph of Section 3.1, we propose a procedure that would allow for the fast adaptation of the system to the users who occur most frequently in the smart home environment and would use such a gender and age classification system. For this classification, we used the notation *Proposed algorithm with segmentation and pitch adaptation*.

Table 2. Accuracy results obtained by gender and age classification of the speaker and statistical analysis of the results with a 95% confidence interval.

Gender and Age Classification of the Speaker	Acc [%]	Mean [%]	95% CI [min max]
(a) Pitch value counts	78.02	78.00	[77.33 78.66]
(b) Pitch value sums	80.98	80.99	[80.32 81.65]
(c) Three-state monophone HMMs—MFCC_Pitch_E_D_A	90.44	90.44	[89.94 90.95]
(d) Sixteen-state word HMMs—MFCC_Pitch_E_D_A	87.18	87.19	[86.63 87.70]
(e) Universal modeling based on GMM clustering [22]	93.32	93.31	[92.88 93.74]
(f) GMMs with segmentation—MFCC_E_D_A	89.32	89.31	[88.77 89.86]
(g) GMMs with segmentation—MFCC_Pitch_E_D_A	91.38	91.39	[90.88 91.85]
(h) Proposed algorithm with segmentation	91.46	91.47	[91.00 91.89]
(i) Proposed algorithm with seg. and pitch adaptation	92.25	92.25	[91.79 92.69]

Table 3. Confusion matrix results obtained by gender and age classification with recall results.

	Male	Female	Child	Recall [%]
(a) Pitch value counts				
Male	**4217**	94	0	**97.8**
Female	514	**3034**	841	**69.1**
Child	185	1124	**2538**	**66.0**
(b) Pitch value sums				
Male	**4262**	49	0	**98.9**
Female	142	**3571**	676	**81.4**
Child	41	1479	**2327**	**60.5**
(c) Three-state monophone HMMs—MFCC_Pitch_E_D_A				
Male	**4270**	31	10	**99.0**
Female	52	**4010**	327	**91.4**
Child	0	780	**3067**	**79.7**
(d) Sixteen-state word HMMs—MFCC_Pitch_E_D_A				
Male	**4282**	19	10	**99.3**
Female	58	**3136**	1195	**71.5**
Child	3	323	**3521**	**91.5**
(e) Universal modeling based on GMM clustering [22]				
Male	**4279**	23	9	**99.3**
Female	76	**4080**	233	**93.0**
Child	12	485	**3350**	**87.1**
(f) GMMs with segmentation—MFCC_E_D_A				
Male	**4108**	159	44	**95.3**
Female	190	**3811**	388	**86.8**
Child	24	535	**3288**	**85.5**
(g) GMMs with segmentation—MFCC_Pitch_E_D_A				
Male	**4264**	43	4	**98.9**
Female	88	**4014**	287	**91.5**
Child	2	657	**3188**	**82.9**
(h) Proposed algorithm with segmentation				
Male	**4271**	36	4	**99.1**
Female	91	**4026**	272	**91.7**
Child	3	665	**3179**	**82.6**
(i) Proposed algorithm with segmentation and pitch adaptation				
Male	**4276**	29	6	**99.2**
Female	62	**4085**	242	**93.1**
Child	5	628	**3214**	**83.5**

The accuracy, *Acc*, presented in Table 2 is defined in (8), where H is the sum of all correct classifications for male, female, and child speakers, divided by the number of all classifications, N.

$$Acc = \frac{H}{N} \cdot 100[\%] \tag{8}$$

The correct classifications for male, female, and child speakers are marked in bold as integer values in Table 3. In Table 3, the results are provided with a different evaluation metric called *Recall*, which is defined in (9), where H_R is the number of correct classifications for male, female, or child speakers in the row divided by the number of all classifications, N_R, in the row for the corresponding class.

$$Recall = \frac{H_R}{N_R} \cdot 100[\%] \tag{9}$$

The number of all classifications, N_R, in the row for the male speakers is 4,311, for the female speakers is 4,389, and for the child speakers is 3,847. From the confusion matrices in Table 3, the calculated *Recall* value can be seen easily. For each classification, the integer value in the row marked in bold is divided by the number of all possible classifications, N_R, in the row for a particular class.

5. Discussion

In this section, we will comment on the results of the experiments in the previous section. First, we will describe the results in Table 2. In addition to the accuracy results, the statistical analysis results with the given confidence interval are also provided. Bootstrapping with 1000 replications was performed for each experiment. As we can see, there were no significant differences between the mean values and the accuracy of the defined test set. If we compare the proposed algorithm with segmentation (experiment h) and the proposed algorithm with segmentation and pitch adaptation (experiment i), it can be seen that the first's accuracy was outside of the second's confidence interval. Thus, we can conclude that the obtained results were statistically significant and not due to chance in the selected test set of the TIDIGITS database [41]. The experiments presented in Table 3 under (a) and (b) mainly obtained their information from the pitch value when determining the speaker's gender and age classification from the recordings. The confusion matrices' results show that, in both cases, the male speaker in the recording was never incorrectly classified as a child speaker. However, the maximum number of confusions in both experiments was present when a female speaker was classified in the recordings even though there was actually a child speaker in the recording. These gender and age classification errors were derived from the pitch values in the training and test sets of the speech database itself, the values of which are presented in Figure 6. The pitch values were determined on the basis of the entire audio recording and the pitch value, which was in the majority of frames in the audio recording, as presented in Figure 6. The pitch values in the children's test set material (Figure 6c) overlapped more severely with the pitch values in the training set material of the female speakers (Figure 6b). Figure 6c shows that the pitch values in the test set were more diffused than in the training set. Experiments (c) and (d) were both based on HMMs. The first used three-state monophone HMMs, and the second used sixteen-state word HMMs. The sixteen-state word HMMs provided worse accuracy (Table 2), most likely due to the use of pitch value as a coefficient in the feature vector. This conclusion is based on the fact that pitch values could only be defined in the voiced speech segment of the word and not through the duration of the whole word, where there were also consonants. Experiment (e) was carried out according to instructions provided elsewhere [22]. Here, the best accuracy was achieved (Table 2), and the recall classification was very good (Table 3)—especially for the child speakers, although the classification of a child speaker was still the most problematic. Such good results were based on more advanced modeling techniques and the use of a larger number of features, such as MPEG-7 low-level descriptors (LLDs), as presented elsewhere [21], and perceptual wavelet packets (PWPs), as presented elsewhere [43]. Our

motivation in this paper was to get closer to these results using less complex procedures that would be more suitable for embedded systems in smart home environments. In the following four experiments—(f)–(i)—we used segmentation based on the proposed VAD algorithm. Experiment (f) was taken as a baseline since it used only MFCC_E_D_A features without pitch values. Comparing experiments (f) and (g), Table 2 shows a more than 2% accuracy improvement in classification performance when the pitch value was used as an additional feature. In the subsequent two experiments—(h) and (i)—an additional contribution can be seen, as we used the classification split into two steps, as proposed in Section 3.1. A more significant contribution was made by the last experiment, (i), when we adapted pitch values. With this experiment, we wanted to present the procedure by which the classification system can be quickly adapted to the normalized pitch values of the users in a smart home environment. As can be seen from experiment (i) in Table 2, the accuracy was better when we performed the adaptation of the proposed classification algorithm with a new set of normalized pitch values. The results in Table 3 show that the recall values of the female speakers for this experiment were the highest of all experiments. After reviewing all the experiments provided in Tables 2 and 3, we can conclude that the biggest problems lie in the classification of female and child speakers. In the classification of male speakers, it is possible to achieve very good results, the values of which were above 99%.

The presented gender and age classification solution can be derived via an embedded system or as a microphone array that captures the signal, and processing is completed on a server. From the user's perspective, how accurately users can be detected by such systems is important. It is required that the system classify the gender and age of the user correctly as often as possible. The speech database used in the tests included 163 speakers, most of whom pronounced 77 isolated or connected digits. Table 4 presents an analysis of the results wherein four parts were identified. First, we checked the number of speakers in which the gender and age classification of the speaker was correct for all audio recordings. In the second and third parts of the analysis, we checked the number of speakers for whom the gender and age were classified incorrectly in 1 to 10 recordings or classified incorrectly in 11 to 20 recordings. If incorrect gender and age classification occurred, the requirement was that the number of these errors be as small as possible. In the fourth and final part of the analysis, we checked the number of speakers in which the gender and age of the speaker were classified incorrectly in 21 to 77 recordings. In this case, in most tests, there were 13 to 17 speakers for which the gender and age were classified incorrectly. When we analyzed these 17 speakers, we found that 11 of them appeared in all tests. There were no male speakers among them, which is also understandable since the male speaker classification was, in most cases, correct. There were eight child speakers and three female speakers for whom gender and age classifications were incorrect in all tests. For these 11 speakers, most confusions in gender and age classification were between the female speaker and child speaker, and vice versa. Of these 11 speakers, 5 speakers were almost entirely incorrectly classified by gender and age in all experiments, which represents 3% of the test material. For these speakers, we can say that they present a challenging task, due to their characteristics, and they will almost always be classified as errors. As can be seen from Table 4, the gender and age classification is presented for all three classes separately (M for male, F for female, and C for child). The column labeled with S represents the sum of values in individual classes. To understand the table better, let us remember that the number of male, female, and child speakers in the test set were 56, 57, and 50, respectively. In experiment (d), for 75 speakers, classification was correct for all audio recordings of these speakers. This represents the best result, but this experiment has as many as 31 speakers for which the gender and age of the speaker were classified incorrectly in 21 to 77 audio recordings. Good results were achieved in experiment (e) due to the very complex methodology, while the results of the proposed algorithm with segmentation and pitch adaptation (experiment (i)) were very similar. However, if we compare experiments (h) and (i), we can see that the adaptation of the system helped to improve the classification

of the female speaker. The number of female speakers for which the female gender was classified correctly in all recordings increased from 10 to 14.

Table 4. Gender and age classification analysis (M—male, F—female, C—child, S—sum of all gender and age classifications) for all 163 speakers.

The Number of Speakers for Which the Gender and Age of the Speaker Were Classified:	**Correctly in All Recordings**				**Incorrectly in 1 to 10 Recordings**				**Incorrectly in 11 to 20 Recordings**				**Incorrectly in 21 to 77 Recordings**			
Experiments: Gender and age:	M	F	C	S	M	F	C	S	M	F	C	S	M	F	C	S
(a) Pitch value counts	49	0	4	53	4	16	19	39	1	17	6	24	2	24	21	47
(b) Pitch value sums	50	11	1	62	4	26	16	46	0	6	10	16	2	14	23	39
(c) Three-state monophone HMMs—MFCC_Pitch_E_D_A	46	9	9	64	9	37	21	67	1	7	9	17	0	4	11	15
(d) Sixteen-state word HMMs—MFCC_Pitch_E_D_A	49	2	24	75	6	11	18	35	1	20	1	22	0	24	7	31
(e) Universal modeling based on GMM clustering [22]	46	12	14	72	10	33	24	67	0	6	4	10	0	6	8	14
(f) GMMs with segmentation—MFCC_E_D_A	30	6	16	52	20	30	19	69	4	11	5	20	2	10	10	22
(g) GMMs with segmentation—MFCC_Pitch_E_D_A	44	11	10	65	11	30	21	62	1	10	8	19	0	6	11	17
(h) Proposed algorithm with segmentation	45	10	12	67	10	31	19	60	1	11	9	21	0	5	10	15
(i) Proposed alg. with segmentation and pitch adaptation	46	14	11	71	10	35	23	68	0	6	4	10	0	2	12	14

The direct comparison of achieved results with other combined age and gender acoustic classification systems is difficult, as experiments were not conducted on the same speech databases (type and amount of speech, language), and also the evaluation conditions differed. A general comparison for the male, female, and child speaker classifications shows that accuracy in previous work [29] was 90.39%, while the proposed system achieved 92.25%, which is a statistically significant improvement. The gap in results between adults and child classes is, in the case of [29], as high as ~35% (classification success: child (60.96%) compared to female (94.50%) or male (96.09%)), while there is a gap in the case of the proposed system i) of between ~10% and ~15% (recall: child (83.5%) compared to female (93.1%) or male (99.2%)). It can be concluded that the inclusion of pitch values improved the classification modeling balance between the child and adult categories.

We performed additional analyses because we wanted to find out how many digits were in the recording when there was an error in the speaker's gender and age classification. Table 5 shows the average number of digits in the audio recordings when the gender and age of the speaker were classified incorrectly. If we compare Tables 4 and 5, we can see that, for 163 speakers, in most cases, errors occurred in 1 to 10 recordings. The number of digits that appeared in these incorrectly classified recordings was, on average, 1.64. In other words, 1 to 2 digits were pronounced in these recordings. The conclusion is that the maximum number of errors occurs in recordings that have small speech content.

Table 5. The average number of digits in the recordings when the gender and age of the speaker were classified incorrectly.

The gender and age of the speaker were classified incorrectly in 1 to 10 recordings.	1.63
The gender and age of the speaker were classified incorrectly in 11 to 20 recordings.	2.38
The gender and age of the speaker were classified incorrectly in 21 to 77 recordings.	3.12

The use of such gender and age speaker classification, based on the acoustic detection of the speaker's presence in the room, is intended primarily for use in intelligent environments of smart homes. The objective is to adapt and personalize services and content to particular user classes. If such a gender and age classification is used in the embedded system, energy consumption is also important. Therefore, we suggest using such a detector, in combination with a PIR motion detection sensor, to turn on the proposed system when a person enters the room. Once gender and age speaker classification, based on acoustic detection, has confirmed the gender or age of the speaker in the room with a high probability, the system can be switched off automatically. Of course, there is still the question of how to act if more people (e.g., an adult and a child) are present in the room. If the acoustic presence detector detects two persons in the room, belonging to different classes (e.g., one is an adult and the other is a child), it is not the task of the acoustic presence detector to define how the smart home environment should react in this case. The issue is resolved at a higher level of the decision support system in the smart home environment, which is not part of the focus of this paper.

6. Conclusions

The presented acoustic presence detection system can be applied in a smart home environment, either as a stand-alone solution or in combination with a PIR motion detection sensor. In this paper, we presented a method for gender and age classification of the speaker, which is based on three different methods completed in two steps. In the first step, the gender and age classification is based on GMMs. Basically, it counts the frames and calculates which frame has a greater probability of belonging to one of the gender and age (male, female, or child) classifications. If the difference between the highest two counted frames belonging to a male, female, or child speaker is less than 4, the second step of gender and age classification is performed, whereby two additional gender and age classification procedures are carried out. The first is based on the count of the frames with normalized pitch values, and the second is based on the sum of the frames with normalized pitch values, which belong to one of the speakers. If all three, or at least two, of the decisions match, then we choose the gender and age that is in the majority. However, if all three decisions are different, we adopt the classification based on GMMs.

Comparative experiments carried out in this paper have shown that algorithms with a large number of different features (some of which are also computationally complex) and more advanced modeling techniques provide slightly better results than in the presented gender and age classification algorithm. However, the proposed classification algorithm was developed for use in a smart home environment, where only simple and efficient classification algorithms are acceptable.

When analyzing the results, we came to the important conclusion that most of the incorrect classifications of the speaker's gender and age occurred in cases where we had a small amount of speech material to analyze. This was, in our case, when only one or two words were captured in the audio recording. We also proposed a procedure that allows us to quickly adapt the gender and age classification algorithm to the frequent users of such a system in smart home environments. The proposed adaptation procedure further improved the performance of speaker gender and age classifications. After performing an extensive set of experiments, we can infer that the proposed method for gender and age classification with adaptations to users could be a potential candidate for integration into real-life smart home environments.

In future work, we will further improve the classification accuracy for children's speech using other low-complexity feature extraction methods, as we aim to use embedded systems for smart home environments.

Author Contributions: Conceptualization, D.V. and A.Z.; methodology, D.V. and A.Z.; validation, D.V. and A.Z.; resources, D.V. and A.Z.; writing—original draft preparation, D.V. and A.Z.; writing—review and editing, D.V. and A.Z.; visualization, D.V.; supervision, A.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Slovenian Research Agency, under Contract No. P2-0069, Research Program "Advanced Methods of Interaction in Telecommunication".

Data Availability Statement: The TIDIGITS speech database is available via LDC: https://doi.org/10.35111/72xz-6x59 (accessed on 22 November 2022).

Acknowledgments: The authors thank Stavros Ntalampiras, for providing us with the source code of the Perceptual Wavelet Packets algorithm. The authors also thank Mirjam Sepesy Maučec, for her valuable comments and suggestions during the writing of this paper.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. United Nations. *World Population to 2300*; Department of Economic and Social Affairs, Population Division: New York, NY, USA, 2004.
2. Mukhamediev, R.I.; Popova, Y.; Kuchin, Y.; Zaitseva, E.; Kalimoldayev, A.; Symagulov, A.; Levashenko, V.; Abdoldina, F.; Gopejenko, V.; Yakunin, K.; et al. Review of Artificial Intelligence and Machine Learning Technologies: Classification, Restrictions, Opportunities and Challenges. *Mathematics* **2022**, *10*, 2552. [CrossRef]
3. Astapov, S.; Gusev, A.; Volkova, M.; Logunov, A.; Zaluskaia, V.; Kapranova, V.; Timofeeva, E.; Evseeva, E.; Kabarov, V.; Matveev, Y. Application of Fusion of Various Spontaneous Speech Analytics Methods for Improving Far-Field Neural-Based Diarization. *Mathematics* **2021**, *9*, 2998. [CrossRef]
4. Giannoulis, P.; Tsiami, A.; Rodomagoulakis, I.; Katsamanis, A.; Potamianos, G.; Maragos, P. The Athena-RC system for speech activity detection and speaker localization in the DIRHA smart home. In Proceedings of the 2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA), Nancy, France, 12 May 2014; pp. 167–171. [CrossRef]
5. Solaimani, S.; Keijzer-Broers, W.; Bouwman, H. What we do–and don't–know about the Smart Home: An analysis of the Smart Home literature. *Indoor Built Environ.* **2015**, *24*, 370–383. [CrossRef]
6. Koo, J.H.; Cho, S.W.; Baek, N.R.; Lee, Y.W.; Park, K.R. A Survey on Face and Body Based Human Recognition Robust to Image Blurring and Low Illumination. *Mathematics* **2022**, *10*, 1522. [CrossRef]
7. Childers, D.G.; Wu, K.; Bae, K.S.; Hicks, D.M. Automatic recognition of gender by voice. In Proceedings of the ICASSP-88, International Conference on Acoustics, Speech, and Signal Processing, New York, NY, USA, 1 January 1988; pp. 603–604. [CrossRef]
8. Wu, K.; Childers, D.G. Gender recognition from speech. Part I: Coarse analysis. *J. Acoust. Soc. Am.* **1991**, *90*, 1828–1840. [CrossRef]
9. Gurgen, F.S.; Fan, T.; Vonwiller, J. On the Analysis of Phoneme Based Features for Gender Identification with Neural Networks. SST. Available online: https://assta.org/proceedings/sst/SST-94-Vol-I/cache/SST-94-VOL1-Chapter9-p8.pdf (accessed on 25 December 2022).
10. Gauvain, J.L.; Lamel, L. Identification of non-linguistic speech features. In Proceedings of the Workshop Held at Plainsboro, Plainsboro, NJ, USA, 21–24 March 1993. [CrossRef]
11. Li, M.; Han, K.J.; Narayanan, S. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Comput. Speech Lang.* **2013**, *27*, 151–167. [CrossRef]
12. Schuller, B.; Steidl, S.; Batliner, A.; Burkhardt, F.; Devillers, L.; Müller, C.; Narayanan, S. Paralinguistics in speech and language—State-of-the-art and the challenge. *Comput. Speech Lang.* **2013**, *27*, 4–39. [CrossRef]
13. Gaikwad, S.; Gawali, B.; Mehrotra, S.C. Gender Identification Using SVM with Combination of MFCC. *Adv. Comput. Res.* **2012**, *4*, 69–73.
14. Yücesoy, E.; Nabiyev, V.V. Gender identification of a speaker using MFCC and GMM. In Proceedings of the 2013 8th International Conference on Electrical and Electronics Engineering (ELECO), Bursa, Turkey, 28 November 2013; pp. 626–629. [CrossRef]
15. Zeng, Y.M.; Wu, Z.Y.; Falk, T.; Chan, W.Y. Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech. In Proceedings of the 2006 International Conference on Machine Learning and Cybernetics, Dalian, China, 13–16 August 2006; pp. 3376–3379. [CrossRef]
16. Müller, C. Automatic Recognition of Speakers' Age and Gender on the Basis of Empirical Studies. In Proceedings of the Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006; Available online: https://www.isca-speech.org/archive/interspeech_2006/muller06_interspeech.html (accessed on 25 December 2022).
17. Islam, M.A. GFCC-based robust gender detection. In Proceedings of the 2016 International Conference on Innovations in Science, Engineering and Technology (ICISET), Dhaka, Bangladesh, 28 October 2016; pp. 1–4. [CrossRef]
18. Meinedo, H.; Trancoso, I. Age and gender detection in the I-DASH project. *ACM Trans. Speech Lang. Process. (TSLP)* **2011**, *7*, 1–6. [CrossRef]
19. Ranjan, S.; Liu, G.; Hansen, J.H. An i-vector plda based gender identification approach for severely distorted and multilingual darpa rats data. In Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, AZ, USA, 13–17 December 2015; pp. 331–337. [CrossRef]

20. Bhavana, R.J.; Swati, P.; Mayur, A. Identification of Age and Gender Using HMM. *Int. J. Comput. Sci. Inf. Technol.* **2015**, *6*, 1643–1647.
21. Ntalampiras, S.; Potamitis, I.; Fakotakis, N. Acoustic detection of human activities in natural environments. *J. Audio Eng. Soc.* **2012**, *60*, 686–695.
22. Ntalampiras, S. A novel holistic modeling approach for generalized sound recognition. *IEEE Signal Process. Lett.* **2013**, *20*, 185–188. [CrossRef]
23. Bocklet, T.; Maier, A.; Bauer, J.G.; Burkhardt, F.; Noth, E. Age and gender recognition for telephone applications based on gmm supervectors and support vector machines. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 1605–1608. [CrossRef]
24. Dehak, N.; Kenny, P.J.; Dehak, R.; Dumouchel, P.; Ouellet, P. Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *19*, 788–798. [CrossRef]
25. Abumallouh, A.; Qawaqneh, Z.; Barkana, B.D. Deep neural network combined posteriors for speakers' age and gender classification. In Proceedings of the 2016 Annual Connecticut Conference on Industrial Electronics, Technology & Automation (CT-IETA), Bridgeport, CT, USA, 14–15 October 2016; pp. 1–5. [CrossRef]
26. Qawaqneh, Z.; Mallouh, A.A.; Barkana, B.D. Deep neural network framework and transformed MFCCs for speaker's age and gender classification. *Knowledge-Based Syst.* **2017**, *115*, 5–14. [CrossRef]
27. Costantini, G.; Parada-Cabaleiro, E.; Casali, D.; Cesarini, V. The Emotion Probe: On the Universality of Cross-Linguistic and Cross-Gender Speech Emotion Recognition via Machine Learning. *Sensors* **2022**, *22*, 2461. [CrossRef]
28. Schuller, B.; Steidl, S.; Batliner, A.; Burkhardt, F.; Devillers, L.; Müller, C.; Narayanan, S. The INTERSPEECH 2010 paralinguistic challenge. In Proceedings of the INTERSPEECH 2010, Makuhari, Japan, 26–30 September 2010; pp. 2794–2797.
29. Yücesoy, E.; Nabiyev, V.V. A new approach with score-level fusion for the classification of a speaker age and gender. *Comput. Electr. Eng.* **2016**, *53*, 29–39. [CrossRef]
30. Meinedo, H.; Trancoso, I. Age and gender classification using fusion of acoustic and prosodic features. In Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, Makuhari, Japan, 26–30 September 2010.
31. Bisio, I.; Delfino, A.; Lavagetto, F.; Marchese, M.; Sciarrone, A. Gender-driven emotion recognition through speech signals for ambient intelligence applications. *IEEE Trans. Emerg. Top. Comput.* **2013**, *1*, 244–257. [CrossRef]
32. Brutti, A.; Ravanelli, M.; Omologo, M. Saslodom: Speech activity detection and speaker localization in domestic environments. In *SASLODOM: Speech Activity Detection and Speaker LOcalization in DOMestic Environments*; Fondazione Bruno Kessler: Povo, Italy, 2014; pp. 139–146.
33. Guerrieri, A.; Braccili, E.; Sgrò, F.; Meldolesi, G.N. Gender Identification in a Two-Level Hierarchical Speech Emotion Recognition System for an Italian Social Robot. *Sensors* **2022**, *22*, 1714. [CrossRef]
34. Tursunov, A.; Choeh, J.Y.; Kwon, S. Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. *Sensors* **2021**, *21*, 5892. [CrossRef]
35. Kwasny, D.; Hemmerling, D. Gender and age estimation methods based on speech using deep neural networks. *Sensors* **2021**, *21*, 4785. [CrossRef]
36. Vlaj, D.; Žgank, A.; Kos, M. Effective Pitch Value Detection in Noisy Intelligent Environments for Efficient Natural Language Processing. In *Recent Trends in Computational Intelligence*; Sadollah, A., Sinha, T.S., Eds.; IntechOpen: London, UK, 2019. [CrossRef]
37. *ETSI Standard ES 201 108 v1.1.1*; Speech Processing, Transmission and Quality aspects (STQ), Distributed Speech Recognition, Front-End Feature Extraction Algorithm, Compression Algorithm. ETSI: Valbonne, France, 2000.
38. Gender and Age Classification Source Code. Available online: https://github.com/dvlaj/FeatureGenderAgeClassification (accessed on 14 December 2022).
39. Anderson, S.R.; Lightfoot, D.W. Describing linguistic knowledge. In *The Language Organ: Linguistics as Cognitive Physiology*; Cambridge University Press: Cambridge, UK, 2002; pp. 92–110.
40. Young, S.; Evermann, G.; Gales, M.; Hain, T.; Kershaw, D.; Liu, X.; Moore, G.; Odell, J.; Ollason, D.; Povey, D.; et al. *The HTK Book*; Cambridge University Engineering Department: Cambridge, UK, 2002; Volume 3, p. 12.
41. Leonard, R. A database for speaker-independent digit recognition. In Proceedings of the ICASSP'84, IEEE International Conference on Acoustics, Speech, and Signal Processing, San Diego, CA, USA, 19 March 1984; Volume 9, pp. 328–331. [CrossRef]
42. Hirsch, H.G.; Pearce, D. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In Proceedings of theASR2000-Automatic Speech Recognition: Challenges for the New Millenium ISCA Tutorial and Research Workshop (ITRW), Paris, France, 18–20 September 2000.
43. Ntalampiras, S.; Potamitis, I.; Fakotakis, N. Exploiting temporal feature integration for generalized sound recognition. *EURASIP J. Adv. Signal Process.* **2009**, *2009*, 807162. [CrossRef]