symmetry

MDPI

Article

Gender Classification Based on the Non-Lexical Cues of Emergency Calls with Recurrent Neural Networks (RNN)

Guiyoung Son [1], Soonil Kwon [1,*] and Neungsoo Park [2]

[1] Department of Software, Sejong University, 209, Neung-dong-ro, Gwangjin-gu, Seoul 05006, Korea; sgy1017@sejong.ac.kr

[2] Department of Computer Science and Engineering Kunkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea; neungsoo@kunkuk.ac.kr

* Correspondence: skwon@sejong.edu

Received: 16 March 2019; Accepted: 10 April 2019; Published: 11 April 2019

Abstract: Automatic gender classification in speech is a challenging research field with a wide range of applications in HCI (human-computer interaction). A couple of decades of research have shown promising results, but there is still a need for improvement. Until now, gender classification has been made using differences in the spectral characteristics of males and females. We assumed that a neutral margin exists between the male and female spectral range. This margin causes misclassification of gender. To address this limitation, we studied three non-lexical speech features (fillers, overlapping, and lengthening). From the statistical analysis, we found that overlapping and lengthening are effective in gender classification. Next, we performed gender classification using overlapping, lengthening, and the baseline acoustic feature, Mel Frequency Cepstral Coefficient (MFCC). We have tried to achieve the best results by using various combinations of features at the same time or sequentially. We used two types of machine-learning methods, support vector machine (SVM) and recurrent neural networks (RNN), to classify the gender. We achieved 89.61% with RNN using a feature set including MFCC, overlapping, and lengthening at the same time. Also, we have reclassified using non-lexical features with only data belonging to the neutral margin which was empirically selected based on the result of gender classification with only MFCC. As a result, we determined that the accuracy of classification with RNN using lengthening was 1.83% better than when MFCC alone was used. We concluded that new speech features could be effective in improving gender classification through a behavioral approach, notably including emergency calls.

Keywords: human-computer interaction; speech processing; emergency call center; gender classification; recurrent neural networks (RNN)

1. Introduction

It is difficult to identify the age, intention, emotion, and gender of a speaker from telephone calls [1]. However, this information is considered essential for automatic speech recognition (ASR) since the cues can guide human-computer interaction systems to understand the needs of users [2,3]. Gender classification is especially useful in the field of ASR because specific acoustic models are applied for the process, which has been reported to improve performance [4]. Furthermore, these can be used in many fields, such as categorizing calls by gender (e.g., for surveys) [4,5]. The systems for gender classification have been recently developed [6,7]. During emergency calls, in particular, it is instrumental in detecting the gender of the caller from the beginning of the call so that the call can be routed to the appropriate receiver according to the caller's gender in order to calm the caller if necessary.

Moreover, the emotional speech recognition systems that considered gender information proved to be more accurate than the system that did not consider it [8]. Therefore, gender classification in emergency calls can be utilized before the recognition systems, i.e., emotion, age, to improve the quality of the interaction. However, gender classification in emergency calls has rarely been studied.

In the ASR, it can accurately classify gender using acoustic information such as fundamental frequency (F0) and Mel Frequency Cepstral Coefficient (MFCC) [9–11]. Frequently, gender classification was based on the pitch as a feature which was presented by References [9,12]. Males have a lower F0 than females during spontaneous speech, vocalization, and text recitation [13,14]. Additionally, males have greater fluctuations in vocal intensity compared to females during spontaneous speech. These results indicate that the F0 of speech is affected by the condition for males whereas females displayed a change in vocal intensity according to the condition. Research on determining differences in verbal expression through jitter, shimmer, MFCC, and other voice features is still ongoing [4,6,15]. However, the above studies have been conducted with voice data of ordinary situations rather than emergency situations. If callers in an emergency are excited or afraid, they may show a different pitch, F0, or MFCC patterns than usual. We need to investigate further to see if it can show the same performance in emergency situations.

So far, MFCC has shown good results in gender classification. However, there has been a limit to the differentiation between a male with a female spectral band and female with a male spectral band. To overcome this, we attempted to find non-lexical speech features that affect gender classification using data from emergency calls. The factors in non-lexical speech utterances during emergency calls that enable gender classification are fillers [16–20], overlapping [16,21–24], and lengthening [24,25].

Fillers are the most common non-lexical speech utterance in a spontaneous speech [18]. They are similar to exclamations, which are very situation dependent, yet they lack any emotional connotations. According to a previous study, mentioned that females frequently use '아' (uh) while males '그' (ku) as fillers [16]. Another study revealed that males more often use '아' (ah), while females showed a tendency to use '어' (uh) as fillers. The sound '어' (ah) used by males has an assertive and extroverted speech utterance characteristic whereas females tend to use '어' (uh) to present a more modest and passive tone in their speech style [20]. Additionally, females use words such as '아' (uh), '음' (um), and '아' (ah) 1.5 to 1.7 times more than males for emphasis [19].

Overlapping occurs when the speaker interrupts another person mid-sentence to express his or her thoughts [23]. In a telephone conversation, males overlap more often than females; when talking to the same gender, females are more likely to make an overlapping with positive expressions for confirmation or agreement. However, males are more likely to make overlapping with the intent to express criticism or make a negative remark. In other words, females tend to show a listening attitude, while males tend to present a selfish attitude when overlapping with another speaker [24]. For this reason, overlapping is a valuable factor for gender classification.

Lastly, lengthening is the formation of a long sound from the last syllable between clauses or phrases during speech and is highly situational. As an agglutinative language, Korean frequently displays combined morphemes in endings as a phrase or sentence (e.g., 내가 집에 가고 있는데 when I am going home). Lengthening is usually presented as an extension using morphemes because the Korean language has an agglutinative nature in general. Especially, lengthening frequently appears in spontaneous speech and reflects the non-fluent speech style of Korean. When communicating in an emergency, lengthening appears to indicate hesitation about the urgency of the current state. In other words, the speaker hesitates because they cannot immediately explain the present situation. It is similar to a kind of blackout that can be caused by excessive emotional expression. In spontaneous speech analysis, females tend to extend the last syllable of the last word in their speech. The average duration of the last syllable is 300 ms, which is quite long [25]. In a speech from travel agents' conversations with customers, there is also a hesitation in the emphasized part of the word, and this can be classified as lengthening [24].

Most of the above papers related to non-lexical features were in the psychological, linguistic, and social fields. The results of those papers depended on only analyzing statistical measures for verification with speech data limited to a normal situation, not an emergency. In addition, unfortunately, they did not attempt any computer scientific approach, nor did they suggest a direction to perform the feature extraction automatically. So, it would be necessary for us to verify the extraction of non-lexical speech utterances more objectively, especially in an emergency.

Recently, most studies analyzing gender classification have focused on machine learning, especially deep learning, using MFCC. In Reference [26], the authors attempted to distinguish gender using AMI meeting corpus. They obtained an accuracy of 90.99% as a result of recurrent neural networks (RNN) with long-short-term memory (LSTM). References [27,28] conducted gender classification using FAU Aibo Emotion Corpus with RNN. It achieved 74.41% in the case of RNN and 76.03% in case of LSTM-RNN. Additionally, Reference [29] described recognizing gender using 22 acoustic parameters on acoustic signals. They achieved 96.74% accuracy on the test dataset.

Further, many studies performed speech recognition using deep learning methods such as artificial neural networks (ANN) and deep neural networks (DNN) for detecting speaker-related information like emotion [30,31], age [27,28], and phoneme [32] across various fields. Recent trends in gender classification have applied multimodal signal processing combined with a variety of data (e.g., text and facial expressions), and using these methods have yielded considerable results [33,34].

The experiments carried out in these papers aimed only to classify using the machine learning or deep learning method. They did not take into consideration a range of specific conditions. Some papers that were cited above used a public database, but did not include emergency situations. AMI meeting corpus [35] are recorded real meetings that include scenario-driven situations, which have been designed to elicit a range of various realistic behaviors. It did not consider the specific condition, such as an emergency situation. FAU Aibo emotion corpus [36] also consisted of communication with a robot named Aibo. Similarly, there was a study which handled telephone speech for annotating voices from normal situatons [37]. It was also applied under normal conditions, not an emergency.

We investigated non-lexical speech utterances in emergency calls. This study was aimed at identifying new features that can help with the gender classification of emergency calls. Primarily, in a previous study, it was difficult to classify gender from the speaker's voice in an unexpected situation (i.e., emergency). In this regard, we need to investigate additional features that can accurately classify gender from the voices in unexpected situations.

We subsequently identified these features and presented the differences between our findings and prior research on non-lexical speech utterances (i.e., fillers, overlapping, lengthening) in emergency calls.

This paper is organized as follows: Section 2 describes the database and the experiment method for gender classification. The results and discussions are presented in Sections 3 and 4, and the conclusions are in Section 5.

2. Materials and Methods

2.1. Materials

The emergency call datasets were taken from the call center of the National Emergency Management Agency (NEIA) of Northern Gyeonggi province. In February 2015, we signed a memorandum of understanding (MOU) with the NEIA of Northern Gyeonggi province to cooperate with the development of technology for social security based on emergency calls.

The voice data were recorded as follows. First, a Zi-Log recording device (VOIP) was installed on four telephones at an emergency call center. This method allowed for the automatic separation of the speaker and the receiver during calls. This voice data were automatically saved as a wave file format. Once the data were recorded, all personal information such as the name and the phone number was deleted. Finally, the data were compressed using the Encryto tool before being saved onto a secure

hard drive (WD Drive 1Tb) and delivered to us. The data in this study were collected with a sampling rate of 8000 Hz, which means the spectral analysis was limited to the range of 0 to 4000 Hz.

We analyzed 335 datasets (male: 225, female: 117) consisting of 342 anonymous Korean native speakers. The speech data length was 22,201 sec (males: 15,360 sec, females: 6841 sec). Seven sets of voice data were excluded. In one set, the first speaker handed the call to a second person, so there were two speakers on the call. In another set, the speaker was a non-Korean who spoke Korean as a second language. In the other two sets of excluded data, no speech was present (Supplementary Materials).

All of the personal information of data had been anonymized by the operator before being transferred to the authors in order to ensure that no contacts should be made with any individuals for the purpose of safeguarding their privacy. In this regard, this study was waived of ethical approval from the Institutional Review Board (IRB) of Sejong University. According to the IRB of Sejong University, research involving the collection or the study of existing data, documents, and records, if these sources are publicly available or if the information is recorded by the investigator in such a manner that the subjects cannot be identified, it can be exempt from IRB review. All the data are anonymous, with no protected personal information included. Therefore, this study meets the requirements for exemption from IRB review.

2.2. Methods

Figure 1 shows an overview of our framework. The input is the combination of 2-channel speech consisting of two speakers. The data were pre-processed to highlight non-lexical speech features which were then verified by statistical analysis. The effective features generated by feature extraction and selection were MFCC, overlapping, and lengthening. Machine learning techniques were utilized to determine the speakers' genders and SVM/RNN were utilized to produce experimental results. The classification system was also computed using majority voting to separate the classified voices from ambiguous voices by a neutral margin. The final output was the percentage likelihood that the voice belonged to a male or female speaker.

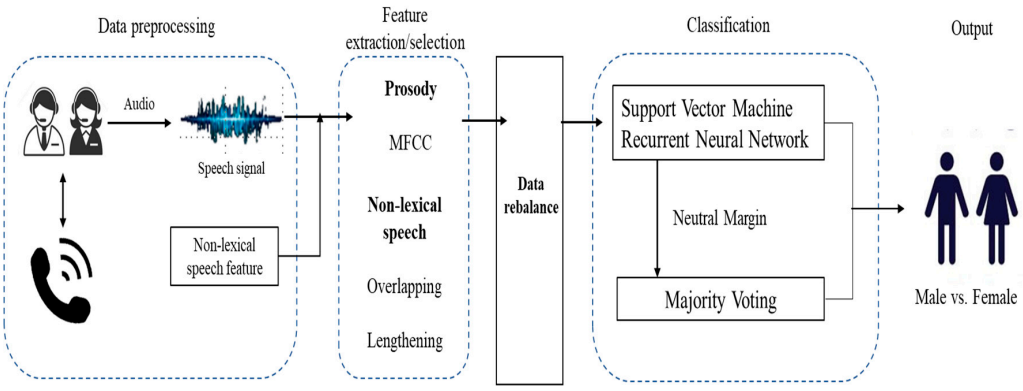

Figure 1. Framework for gender classification using emergency calls.

2.2.1. Non-Lexical Speech Utterances for Gender Classification

We investigated the occurrence frequency and the number of frequency of the three non-lexical speech utterances (fillers, overlapping, lengthening) for each call. To minimize errors in manual gender classification by a group of listeners, we conducted crosschecks for each call. The listeners also practiced beforehand with training data before analysis.

We investigated differences in non-lexical speech utterances by gender, with a focus on the following: The occurrence frequency, the number of occurrences (if occurred), and the results of correlation among non-lexical speech features based on statistical analysis. Based on previous research [16,18,20], the five most common filler words were identified as follows: '어' (ah), '그' (ku), '저' (ceo), '아' (uh), and '음' (um) (Table 1).

Table 1. The example of fillers.

Filler	Examlple
아 (ah)	***아*** 아니 (***ah*** A Ni) = ***ah*** No
어 (uh)	***어 어*** 있잖아 (***uh uh*** I chana) = **uh uh** you know
음 (um)	내가 ***음*** 많이 아파요. (Neyga ***um*** Mani Appayo.) = I am ***um*** very sick.
저 (ceo)	***저*** 여기 00 식당인데요. (***Ceo*** Yeogi 00 Siktangindeyo.) = ***ceo*** this is 00 restaurant.
그 (ku)	***그*** 여기가 ***그*** (***ku*** Yeogiga ***ku***) = ***ku*** here is ***ku***

Overlapping was defined when the speaker started to speak before the receiver had finished talking. Every instance in which the speaker interrupted the receiver was counted as one occurrence.

Lengthening was defined as the phenomenon of prolonged sounds between clauses and phrases uttered by the speaker, usually occurring at the end of words or phrases. Based on prior research [25], lengthening was identified when spoken syllables were prolonged for more than 300 ms.

2.2.2. Statistical Analysis

We used the following methods to find the differences in non-lexical speech utterances for gender classification using statistical analysis: Pearson's Chi-Square test, independent sample test, and Pearson's correlation analysis. It was essential to validate the occurrence frequency and the number of frequency for feature selection. The significance levels were set at $p < 0.05$ and analysis was conducted with SPSS 21.0 (IBM).

We created a decision tree to identify useful factors for gender classification. The decision tree is a type of data mining tool used to create a model that will detect correlations and patterns within data [38]. We created classification and regression trees (CRT) in an attempt to maximize within-node homogeneity (Figure 2). CRT is a method used for finding the frequency weights or influential variables that best delineate the dependent variables. Also, we performed 10-fold cross-validation to assess how well our tree structure can be generalized to a larger population. The data used for this process were randomly selected. The tree generated the nodes in sequence, excluding the data from each subsample in turn. The cross-validation produced a single final model. The cross-validated risk estimate for the last nodes was calculated as the average of the risks for all of the trees [38].

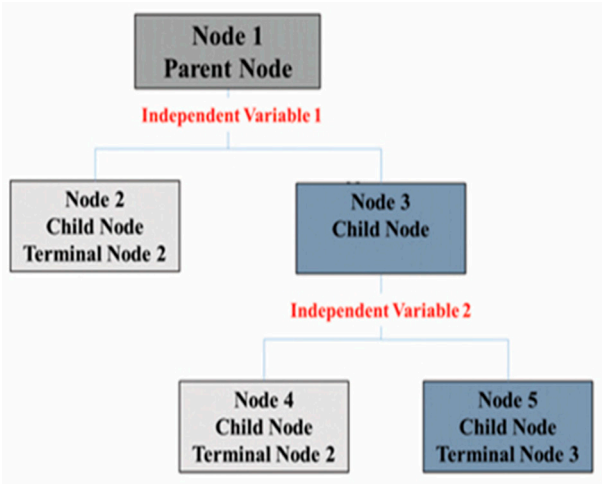

Figure 2. The example of the classification and regression tree (CRT) model.

2.2.3. Machine-Learning-Based Gender Classification

We present a method for gender classification of speech from emergency calls using the machine learning technique known as support vector machines (SVM). SVM is employed for solving two-group classification problems [39] and is used to construct the optimal hyperplane with the largest margin for separating data between two groups. This classifier is widely used for pattern recognition or data analysis. We used Scikit-learn toolbox based on Python [40].

We also utilized RNN, which is a powerful model for processing time-sequential data [41]. RNN was recently demonstrated to outperform feed-forward networks in speech recognition [42]. The RNN architecture is incorporated with Tensorflow using the basic RNN Cell model and consists of one fully connected layer (two dimensions) with 128 (dimensions). The loss function used was sigmoid cross-entropy with logits, and the learning rate was set to 0.0001. Each training epoch consisted of 500 such instances. We used RMS Prop Optimizer for optimization and applied a dropout with 0.5 rates.

We conducted gender classification using three features. The commonly used MFCC feature was considered. MFCC is useful for gender classification. We extracted the initial 16 coefficients produced with 19 filters in the Mel-filter bank. We hypothesized that a neutral margin exists between male and female voice data. Also, the neutral margin causes the misclassification of gender. Hence, other features were used for the two non-lexical speech utterances (i.e., overlapping and lengthening) described in Section 3. Five-fold cross validation was used with 340 sets to test the gender classification results. We have proceeded in two steps to complement the existing method which used only MFCC.

First, we performed gender classification with SVM and RNN using MFCC and non-lexical speech features (overlapping and lengthening) at the same time. The speech data length used for machine learning was 13,676.58 sec. Next, we sequentially reclassified genders using non-lexical speech features to clarify some vague or slightly unclear results in the previous gender classification using MFCC alone. In determining the neutral margin between the two genders, we applied an empirical research method which resulted in a range of 10% above or below the boundary of probability distribution functions of the two genders. The speech data length included in the neutral margin was 1831.14 sec of SVM and 1590.84 sec of RNN.

3. Results

3.1. Descriptive Analysis: Non-Lexical Speech Utterances

Table 2 shows the occurrence frequencies of three non-lexical speech utterances (fillers, the overlapping, lengthening). The occurrence frequencies of overlapping and lengthening were useful for gender classification. Males had a higher occurrence frequency for overlapping than females (146 > 45), but the result was the opposite for lengthening. The difference was significant by gender ($p < 0.001$).

Table 2. The occurrence frequency per minute by non-lexical speech utterances.

Speech Utterance	Occurrence	Gender (N = 340)		x^2	df	p
		Male (N = 223)	Female (N = 117)			
Fillers	Occurrence	197 (88.3)	75 (64.1)	28.177	1	0.001
	Non-occurrence	26 (11.7)	42 (35.9)			
Overlapping	Occurrence	146 (65.5)	45 (38.5)	22.739	1	0.001
	Non-occurrence	77 (34.5)	72 (61.5)			
Lengthening	Occurrence	145 (63.7)	93 (79.5)	8.986	1	0.003
	Non-occurrence	81 (36.3)	24 (20.5)			

Numbers represent the number of subjects (percentages)/$p < 0.05$.

Table 3 shows the mean frequency per minute for the three non-lexical speech utterances. The results were significant for overlapping and lengthening. Overlapping was used more frequently by males than females, but lengthening was the reverse. However, the use of fillers was not significantly different because the usage was high for both genders.

Table 3. The number of mean frequency in the non-lexical speech utterances between genders.

Speech Utterance	Male	Female	t	df	*p*-Value
	Mean (SD)	Mean (SD)			
Fillers	3.004 (2.542)	2.586 (1.807)	1.319	274	0.188
Overlapping	1.851 (1.407)	1.482 (1.421)	2.294	193	0.023
Lengthening	2.493 (2.080)	4.479 (2.901)	−5.711	153.07	0.001

Numbers represent the number of frequency (Standard Deviation)

Table 4 shows the Pearson's coefficient correlation for the identified features. Overlapping had a significant correlation with fillers and lengthening, though the coefficient for lengthening (r = 0.454, $p = 0.000$) was much higher. The results for lengthening were similar to those for overlapping. However, with fillers, the coefficient for overlapping for females (r = 0.483, $p = 0.006$) was higher than that for males (r = 0.249, $p = 0.003$). Additionally, fillers did not have a significant correlation with lengthening (r = 0.098, $p = 0.173$).

Table 4. The correlation between the non-lexical speech utterances (fillers, overlapping, and lengthening).

			Target Variable: Gender					
			Node by Node					
		Node	Node		Response		Gain	Gain Index
			N	%	N	%		
Male	Training	4	104	44.7%	85	56.9%	82.8%	128.8%
		3	69	29.6%	45	29.7%	64.2%	99.9%
		2	60	25.6%	20	13.2%	33.1%	51.6%
	Test	4	28	45.0%	22	54.9%	78.3%	121.9%
		3	20	32.1%	13	30.9%	61.3%	94.9%
		2	14	22.8%	6	14.1%	38.7%	60.3%
Female	Training	2	60	25.6%	40	47.8%	66.8%	187.5%
		3	69	29.6%	25	29.3%	35.7%	100.1%
		4	104	44.7%	19	22.7%	17.1%	48.4%
	Test	2	14	22.8%	9	38.8%	61.2%	175.5%
		3	20	32.1%	7	33.5%	38.6%	107.5%
		4	28	45.0%	6	27.5%	21.6%	61.0%

We selected valuable factors based on the coefficient values and *p*-values for gender classification. We decided on two non-lexical speech utterances, overlapping and lengthening. Next, we created a decision-making tree, which is a tool for data mining analysis, for gender classification.

3.2. Decision-Making Tree: CRT Analysis

We performed 10-fold cross validation using CRT analysis for gender classification based on overlapping and lengthening in emergency calls. Figure 3 shows one example of the structure models developed using CRT analysis. As shown in Figure 3, the cut-off point is vital for gender classification.

The node was divided by lengthening; females were above 3.86 and males were below. The node was split again by a cut-off point for overlapping, specifically a frequency per minute of 1.60.

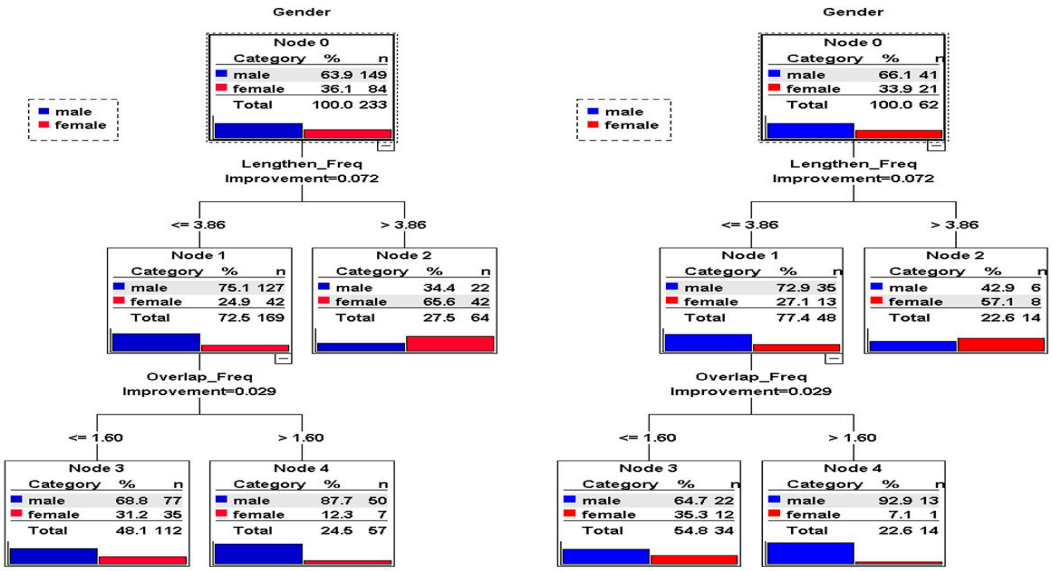

Figure 3. The example of one sample structure model in gender classification using CRT (a 10-fold cross validation); left: Training set, right: Test set.

Table 5 shows the gain chart for the nodes used for gender classification with mean values. For males, the ratio of node 4 in the training set was 44.7% (N = 104) of the total (N = 233), and the percentage ratio of responses was 56.9% (N = 85) for male classification (N = 150). Males had the highest mean value at node 4 while females had a higher mean value at node 2. According to the training datasets, node 4 (N = 85, 82.8%) was the best node for classifying males. Females were placed at node 2 (N = 40, 66.8%) with the training set.

Table 5. Gain chart in nodes to predict gender classification using mean values.

Variable		Total (N = 340)			Male (N = 223)			Female (N = 117)		
		F	**O**	**L**	**F**	**O**	**L**	**F**	**O**	**L**
F	r	1			1			1		
	p									
	n	276			199			77		
O	r	292 **	1		.249 **	1		483 **	1	
	p	.000			.003			.006		
	n	168	195		137	149		31	46	
L	r	.098	.454 **	1	.109	.433 **	1	.241	.555 **	1
	p	.173	.000		.217	.000		.053	.001	
	n	194	135	235	129	101	142	65	34	93

r: coefficient of correlation, *p*: *p*-value, n: sample size; F: Fillers, O: overlapping, L: Lengthening; *: $p < 0.05$, **: $p < 0.01$

As shown in Table 6, the classification accuracy was demonstrated to be 72.9% in the training set and 69.8% in the test set. This value was a result of 69.8% accuracy in the test set because of the training set. In summary, the predicted classification accuracy was approximately 72.9% in the training set and 69.2% in the test set. It means that the predicted accuracy was relatively high.

Table 6. The mean accuracy for gender.

Category		Predicted		
		Male	**Female**	**Accuracy**
Training	Male	130 (6)	20 (3)	86.9%
	Female	40 (1)	44 (2)	47.7%
	Overall percentage	79.7%	20.3%	72.9%
Risk Estimate			.271	
Risk Std. Error			.029	
Test	Male	35 (5)	6 (2)	86.9%
	Female	13 (1)	9 (2)	38.3%
	Overall percentage	82.5%	17.5%	69.8%
Risk Estimate			.308	
Risk Std. Error			.058	

(): Standard deviation, Growing Method: CRT, Dependent Variable: Gender

Based on these results, the summary for gender classification is as follows: (1) The occurrence frequency of overlapping and lengthening is critical for gender classification. (2) The "cut-off points" for overlapping and lengthening [38] are different for each gender and (3) the accuracy of gender classification using overlapping and lengthening is 69.2%.

3.3. Gender Classification: SVM and RNN

Figure 4 shows the gender classification results using baseline acoustic features (MFCC) and non- lexical speech features (overlapping and lengthening). We achieved the highest accuracy when overlapping and lengthening were combined with MFCC, compared to using only MFCC or the other combination of features. The accuracy of classification based on RNN was slightly higher than that of SVM in the case of each feature set. Notably, the highest accuracy was 86.58% of SVM and 89.61% of RNN, separately, when all of the features (MFCC, overlapping, and lengthening) were used simultaneously.

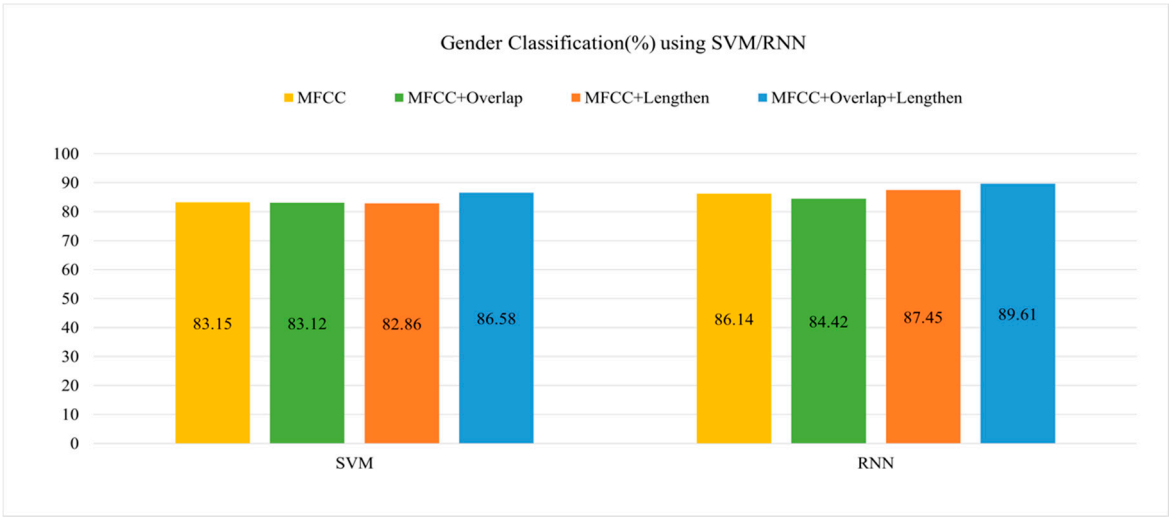

Figure 4. Gender Classification using support vector machine (SVM)/recurrent neural networks (RNN).

Furthermore, we experimented using a neutral margin, defined as above, for investigating what kind of feature set is the most effective as well as how to use it for the improvement of gender

classification. First, we empirically found the neutral margin based on the prior result, which was done with only MFCC. Then, we reclassified data within the neutral margin. With regard to the neutral margin, we achieved a higher accuracy with lengthening than with other non-lexical speech features, including MFCC only. For SVM, it gained about 0.6% better accuracy over MFCC only (Figure 5). For RNN, the accuracy of non-lexical features was significantly improved, especially lengthening (Figure 6) compared to the other non-lexical speech features, as well as MFCC only.

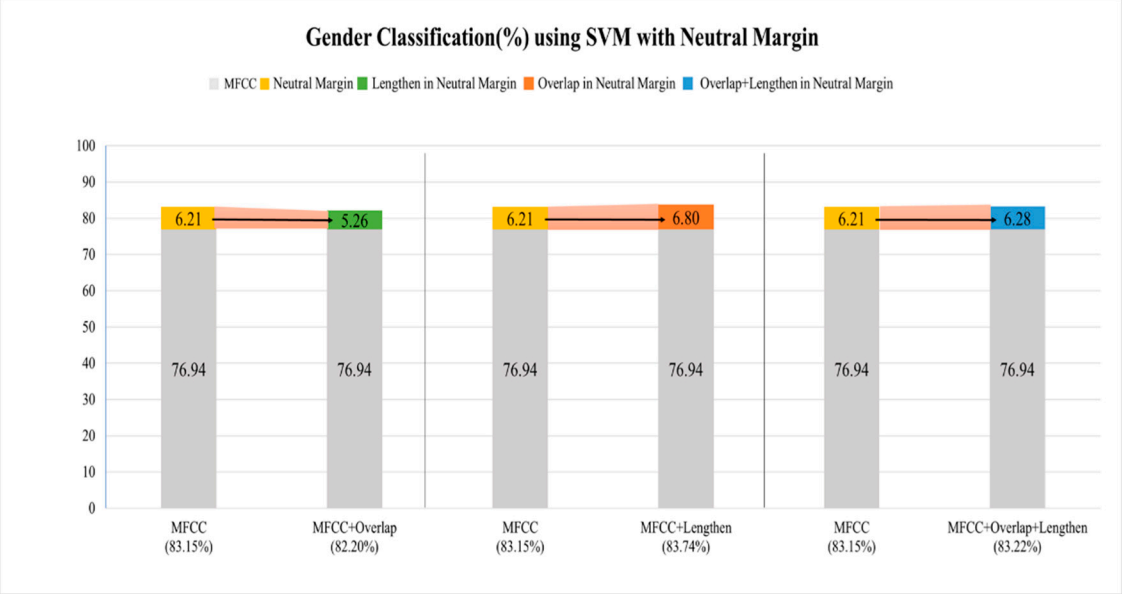

Figure 5. Gender classification using SVM with neutral margin.

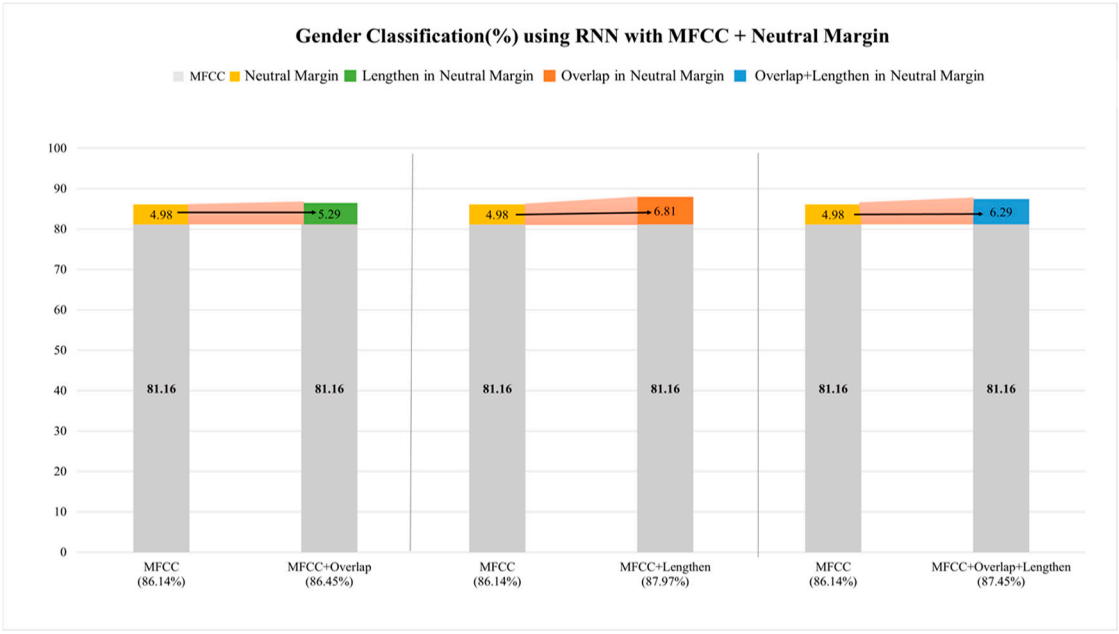

Figure 6. Gender classification using RNN with neutral margin.

The best result was obtained using all of the features at the same time with RNN. Our experiments showed that with all of the data, the combination of overlapping and lengthening achieved a better performance than both MFCC only and overlapping or lengthening separately. However, with the neutral margin, the combination of overlapping and lengthening features gained a higher performance than overlapping and a slightly worse performance than lengthening.

4. Discussion

This experiment was conducted to determine whether the experimental method could accurately and automatically classify the gender of the speaker making a call to an emergency response center. The database targeted for this study should have reflected the actual situation. Over the last few decades, researchers have been trying to find new ways to classify speaker gender. MFCC has commonly been used as the baseline feature because it is based on the typical spectral differences of male and female voices. However, it has difficulty distinguishing the gender of speakers whose voices have intermediate spectral characteristics.

This study was conducted to overcome the shortcomings of existing methods in identifying speaker gender due to the various emotional states expressed in emergencies. This study was predicted to be challenging because the situational factors examined in this study were different from those examined in other studies. If situation factors can be considered, higher accuracy would be obtained by combining non-lexical speech utterances that reflect situational conditions such as emergencies. Thus, to overcome the limitations of existing methods, this study tested analyzed non-verbal, non-lexical behavioral cues that have not been used for gender classification before. Statistical analysis of emergency calls verified that overlapping and lengthening are useful non-lexical speech utterance factors for gender classification.

We also conducted gender classification using the machine-learning techniques, SVM and RNN, with some features (MFCC, overlapping, and lengthening). As expected based on previous analyses, the combination of the proposed non-lexical features and the more traditional MFCC produced more accurate gender classification results than the use of MFCC alone (Figure 4). This result indicated that the proposed features can be used to overcome the shortcomings of existing gender classification methods.

This study was conducted to determine whether the proposed method could overcome the limitations of existing gender identification methods using only MFCC. It was hypothesized that it would be difficult to determine a speaker's gender within the neutral margin. This study investigated the effectiveness of using non-lexical speech features in determining the gender of speakers within the neutral margin. The results showed that non-lexical features were more effective than commonly used spectral features for determining the gender of speakers in the neutral margin. The proposed method, which used MFCC with overlapping and lengthening in the classification of the gender of speakers in neutral margin, was more accurate than using only MFCC. In conclusion, the proposed gender classification method can overcome the weaknesses of existing gender classification methods (Figures 5 and 6). It can be considered both challenging and novel.

Even though gender classification through speech in emergency situations includes more obstacles, such as noise and emotional vocal traits, than in normal ones, the method proposed in this study performed better than standard methods. However, this study had some limitations the future research should account for. It was difficult to use both non-lexical speech features and MFCC over time because the non-lexical speech feature occurred with a different frequency than X over time. For example, it was possible to confirm overlapping by determining when a speaker finished an utterance, but lengthening sometimes appeared while continuing speech. However, MFCC occurs more quickly than non-lexical speech features, so it is difficult to evaluate them when they are combined. If the difference of timescales is more effectively reflected in the experiment's setup, we can expect a higher experimental accuracy. In addition, we have focused only on the speaker's gender without considering the receiver's gender in this study. If we try to classify the gender of both a caller and a receiver in emergency calls, we will likely identify additional speech features and statistic outcomes such as cut-off points or predictability.

5. Conclusions

We investigated the usability of non-lexical speech utterances for gender classification with emergency calls to address the limitation of the baseline feature, MFCC. This study is attractive because of non-lexical speech utterances in special situations, not normal. Furthermore, the non-lexical speech

utterances proposed in this study may be utilized as supporting materials for speech recognition and speech processing.

We also confirmed that overlapping and lengthening had complementary effects when combined with MFCC, indicating that gender classification can be improved by using these two features (acoustic and non-lexical speech features). Furthermore, if we combine both features, we can expect to achieve a higher accuracy in results using machine learning (e.g., SVM, RNN). In summary, the results may assist in classifying the gender of a speaker when analyzing voices in unexpected situations, such as emergencies.

We expect that the results will substantially contribute to the improvement of gender classification for an integrated system for efficient intelligent support systems that can be used for emergency rescues. Furthermore, our findings can help to advance non-lexical speech utterance features extracted from a speaker's voice for situation awareness.

Supplementary Materials: Supplementary Materials are available at: http://www.mdpi.com/2073-8994/11/4/525/s1. The current study used recorded voice data, which can be obtained from the call center of the National Emergency Management Agency (NEIA) of Northern Gyeonggi Province (Korea). Due to ethical restrictions imposed by the Ethics Committee of the Sejong University and the MOU agreements with the call center of the National Emergency Management Agency (NEIA) of Northern Gyeonggi Province (Korea), the data cannot be made publicly available. The interested parties may request the data from the first author: sgy1017@sejong.ac.kr.

Author Contributions: G.S. and S.K. conceived and designed the experiments; G.S. and N.P. investigated the previous works and methodology; all authors performed the experiments; G.S. and S.K. analyzed the data; G.S. wrote the paper; S.K. reviewed and edited; S.K. supervised all of the progress.

Funding: This research work was partly supported by the Institute for Information & Communications Technology Planning & Promotion (IITP), and the grant funded by the Korean government (MSIP) (No. R0126-15-1119. The development of a solution for situation awareness based on the analysis of speech and environmental sounds).

Acknowledgments: This research work was partly supported by the Institute for Information & Communications Technology Planning & Promotion (IITP), and the grant funded by the Korean government (MSIP) (No. R0126-15-1119. The development of a solution for situation awareness based on the analysis of speech and environmental sounds).

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

References

1. Groves, R.M.; O'Hare, B.C.; Gould-Smith, D.; Benki, J.; Maher, P. Telephone interviewer voice characteristics and the survey participation decision. *Adv. Teleph. Surv. Methodol.* **2008**, 385–400.
2. 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]
3. Siniscalchi, S.M.; Salerno, V.M. Adaptation to new microphones using artificial neural networks with trainable activation functions. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *28*, 1959–1965. [CrossRef] [PubMed]
4. Naini, A.S.; Homayounpour, M. Speaker age interval and sex identification based on jitters, shimmers and mean mfcc using supervised and unsupervised discriminative classification methods. In Proceedings of the 2006 8th international Conference on Signal Processing, Beijing, China, 16–20 November 2006.
5. 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.
6. Metze, F.; Ajmera, J.; Englert, R.; Bub, U.; Burkhardt, F.; Stegmann, J.; Muller, C.; Huber, R.; Andrassy, B.; Bauer, J.G. Comparison of four approaches to age and gender recognition for telephone applications. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, Honolulu, HI, USA, 15–20 April 2007; pp. 1089–1092.
7. Vergin, R.; Farhat, A.; O'Shaughnessy, D. Robust gender-dependent acoustic-phonetic modelling in continuous speech recognition based on a new automatic male/female classification. In Proceedings of the Fourth International Conference on Spoken Language Processing, Philadelphia, PA, USA, 3–6 October 1996; pp. 1081–1084.

8. Ververidis, D.; Kotropoulos, C. Automatic speech classification to five emotional states based on gender information. In Proceedings of the EUSIPCO, Vienna, Austria, 6–10 September 2004; pp. 341–344.
9. Hu, Y.; Wu, D.; Nucci, A. Pitch-based gender identification with two-stage classification. *Secur. Commun. Netw.* **2012**, *5*, 211–225. [CrossRef]
10. Ting, H.; Yingchun, Y.; Zhaohui, W. Combining MFCC and pitch to enhance the performance of the gender recognition. In Proceedings of the 2006 8th international Conference on Signal Processing, Beijing, China, 16–20 November 2006.
11. Kabil, S.H.; Muckenhirn, H.; Doss, M.M. On Learning to Identify Genders from Raw Speech Signal using CNNs. 2018. Available online: http://publications.idiap.ch/downloads/papers/2018/Kabil_INTERSPEECH_2018.pdf (accessed on 1 March 2019).
12. Barkana, B.D.; Zhou, J. A new pitch-range based feature set for a speaker's age and gender classification. *Appl. Acoust.* **2015**, *98*, 52–61. [CrossRef]
13. Sapienza, C.M. Aerodynamic and acoustic characteristics of the adult AfricanAmerican voice. *J. Voice* **1997**, *11*, 410–416. [CrossRef]
14. Morris, R.J.; Brown, W.; Hicks, D.M.; Howell, E. Phonational profiles of male trained singers and nonsingers. *J. Voice* **1995**, *9*, 142–148. [CrossRef]
15. Hanson, H.M.; Chuang, E.S. Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *J. Acoust. Soc. Am.* **1999**, *106*, 1064–1077. [CrossRef] [PubMed]
16. Jun, J.; Kim, S. Gender Differences in Powerful/Powerless Language Use in Adult and Higher Education Settings: A Meta-Analysis. *Asian J. Educ.* **2005**, *6*, 53–73.
17. Corley, M.; Stewart, O.W. Hesitation disfluencies in spontaneous speech: The meaning of *um*. *Lang. Linguist. Compass* **2008**, *2*, 589–602. [CrossRef]
18. Stouten, F.; Duchateau, J.; Martens, J.-P.; Wambacq, P. Coping with disfluencies in spontaneous speech recognition: Acoustic detection and linguistic context manipulation. *Speech Commun.* **2006**, *48*, 1590–1606. [CrossRef]
19. Hye-Young, K. A Corpus-analysis of Gender Effects in Private Speech: The Function of Discourse Markers in Spoken Korean. *Lang. Linguist.* **2011**, *53*, 89–108.
20. Gyu-hong, L. A Study on the Use of Korean Discourse Markers according to Gender. *Korean Lang. Lit.* **2004**, *1*, 93–113.
21. Soon-ja, K. Special Feature-Korean Speech and Conversation Analysis: Characteristics found among Men and Women engaging in Interrupting the Turn of the Next Speaker. *Speech Res.* **2000**, *2*, 61–92.
22. Cheon, E.S. The Difference between Men's and Women's Speeches: From the Aspect of Discourse Strategy and Discourse Context. *Korean J. Russ. Lang. Lit.* **2007**, *19*, 41–73.
23. Won-Pyo, L. Interventions in Talk Shows: Discourse Functions and Social Variables. *Discourse Cogn.* **1999**, *6*, 23–59.
24. Kim, S.-H. Intonation Patterns of Korean Spontaneous Speech. *J. Korean Soc. Speech Sci.* **2009**, *1*, 85–94.
25. Min-Ha, J. Politeness Strategy in Intonation Based on Age: Through Analysis of Spontaneous Speech of Those in 10s, 20s, and 30s Women. *Korean Semant.* **2014**, *45*, 99–127.
26. Hagerer, G.; Pandit, V.; Eyben, F.; Schuller, B. Enhancing lstm rnn-based speech overlap detection by artificially mixed data. In Proceedings of the Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio, Erlangen, Germany, 22–24 June 2017; p. 1.
27. Wang, Z.-Q.; Tashev, I. Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. In Proceedings of the 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5150–5154.
28. Buyukyilmaz, M.; Cibikdiken, A.O. Voice gender recognition using deep learning. In Proceedings of the 2016 International Conference on Modeling, Simulation and Optimization Technologies and Applications (MSOTA2016), Xiamen, China, 18–19 December 2016.
29. Bisio, I.; Delfino, A.; Lavagetto, F.; Marchese, M.; Sciarrone, A. Gender-driven emotion recognition through speech signals for ambient intelligence applications. *IEEE Emerg. Top Com.* **2013**, *1*, 244–257. [CrossRef]
30. Zhang, L.; Wang, L.; Dang, J.; Guo, L.; Yu, Q. Gender-Aware CNN-BLSTM for Speech Emotion Recognition. In Proceedings of the International Conference on Artificial Neural Networks, Rhodes, Greece, 4–7 October 2018; pp. 782–790.

31. Zazo, R.; Nidadavolu, P.S.; Chen, N.; Gonzalez-Rodriguez, J.; Dehak, N. Age estimation in short speech utterances based on LSTM recurrent neural networks. *IEEE Access* **2018**, *6*, 22524–22530. [CrossRef]
32. Siniscalchi, S.M.; Yu, D.; Deng, L.; Lee, C.-H. Exploiting deep neural networks for detection-based speech recognition. *Neurocomputing* **2013**, *106*, 148–157. [CrossRef]
33. Katerenchuk, D. Age group classification with speech and metadata multimodality fusion. *arXiv* **2018**, arXiv:1803.00721.
34. Abouelenien, M.; Pérez-Rosas, V.; Mihalcea, R.; Burzo, M. Multimodal gender detection. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; pp. 302–311.
35. McCowan, I.; Carletta, J.; Kraaij, W.; Ashby, S.; Bourban, S.; Flynn, M.; Guillemot, M.; Hain, T.; Kadlec, J.; Karaiskos, V. The AMI meeting corpus. In Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research, Edinburgh, UK, 11–13 July 2005; p. 100.
36. Batliner, A.; Steidl, S.; Nöth, E. Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus. In Proceedings of the Satellite Workshop of LREC, Marrakech, Morocco, 26 May–1 June 2008; pp. 28–31.
37. Burkhardt, F.; Eckert, M.; Johannsen, W.; Stegmann, J. A Database of Age and Gender Annotated Telephone Speech. In Proceedings of the LREC, Valletta, Malta, 17–23 May 2010; pp. 1562–1565.
38. IBM SPSS Decision Trees 21. 2012. Available online: http://www.sussex.ac.uk/its/pdfs/SPSS_Decision_Trees_21.pdf (accessed on 1 March 2019).
39. Cortes, C.; Vapnik, V. Support vector machine. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
40. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
41. Graves, A.; Mohamed, A.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6645–6649.
42. Eyben, F.; Weninger, F.; Squartini, S.; Schuller, B. Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 483–487.