computers

MDPI

Article

An Investigation of a Feature-Level Fusion for Noisy Speech Emotion Recognition

Sara Sekkate [1,*], Mohammed Khalil [1], Abdellah Adib [1] and Sofia Ben Jebara [2]

[1] Team Networks, Telecoms & Multimedia, University of Hassan II Casablanca, Casablanca 20000, Morocco; medkhalil87@gmail.com (M.K.); adibab@gmail.com (A.A.)
[2] COSIM Lab, Higher School of Communications of Tunis, Carthage University, Ariana 2083, Tunisia; sofia.benjebara@supcom.tn
* Correspondence: sarasekkate@gmail.com

Received: 18 October 2019; Accepted: 11 December 2019; Published: 13 December 2019

check for updates

Abstract: Because one of the key issues in improving the performance of Speech Emotion Recognition (SER) systems is the choice of an effective feature representation, most of the research has focused on developing a feature level fusion using a large set of features. In our study, we propose a relatively low-dimensional feature set that combines three features: baseline Mel Frequency Cepstral Coefficients (MFCCs), MFCCs derived from Discrete Wavelet Transform (DWT) sub-band coefficients that are denoted as DMFCC, and pitch based features. Moreover, the performance of the proposed feature extraction method is evaluated in clean conditions and in the presence of several real-world noises. Furthermore, conventional Machine Learning (ML) and Deep Learning (DL) classifiers are employed for comparison. The proposal is tested using speech utterances of both of the Berlin German Emotional Database (EMO-DB) and Interactive Emotional Dyadic Motion Capture (IEMOCAP) speech databases through speaker independent experiments. Experimental results show improvement in speech emotion detection over baselines.

Keywords: speech emotion recognition; feature fusion; SVM; naive Bayes; wavelet

1. Introduction

Speech signals encompass a large set of information, ranging from lexical content to speaker's emotional state and traits. Decoding such information has been of benefit to a number of different speech processing tasks such as speech and speaker recognition, as well as Speech Emotion Recognition (SER). After long-term research, both speech and speaker recognition have been addressed pretty well [1–5], while SER remain difficult, especially in the presence of noise.

Although SER is a widely used research topic, most existing systems are processed under ideal acoustic conditions. Noise is a major factor that affects the performance of speech related tasks [6], but it is still not well studied in the emotional context. Hence, one of the main purposes of this work is to propose an implementation of a SER system in a more realistic setting. It consists of detecting speakers' emotional state under acoustic corruption caused by real-word noises.

A major issue of SER in adverse environment resides in the fact that both background noise and emotion are intermingled together. This implies that whenever we want to extract information about the speaker's emotional state, noise contributes as an uncertainty that results in an acoustic mismatch between training and testing conditions for real-life scenarios. In order to reduce this mismatch, extensive research has been conducted and has tried to intervene in different stages of the recognition process. The developed methods fall into two main categories. The first one includes speech enhancement or noise reduction techniques, either by means of speech sample reconstruction [7], noise compensation using histogram equalization [8], adaptive thresholding in the wavelet domain for noise cancellation [9], or spectral subtraction [10]. However, the downside of such methods is that they strongly rely on prior knowledge of noise, speech, or both, which limits their implementations. The second category includes methods that utilize no prior knowledge of the background noise. They concentrate on finding noise robust features either by proposing new ones such as Log Frequency Power Ratio (LFPR) [11], Rate Scale (RS) [12], Teager Energy based Mel Frequency Cepstral Coefficients (TEMFCCs) [13], Power Normalized Cepstral Coefficients (PNCC) [14], or constructing large feature sets by combining different features.

Since most of the available works concentrate on only a particular type of noise, the contribution of this paper is twofold: on the one hand, it represents a new effort to approach real-life conditions and presents an SER system that is subjected to different types of noise; on the other hand, it proposes a multi-feature fusion representation that is based on a combination of conventional features and wavelet analysis and shows that the usage of a relatively low-dimensional feature set can be conceivable for application to SER even in unconstrained conditions. The organization of the paper is as follows: In Section 2, we provide an extensive overview of various SER works that are based on feature-level fusion. Next, we introduce in Section 3 the proposed framework and the proposed emotion recognition system. In Section 4, we evaluate the performances of the proposed work using two corpora. Finally, Section 5 highlights the contributions of the present work and provides potential future research directions.

2. Review of Feature Level Fusion Based SER

One of the most important consideration for SER is the extraction of the suitable features, which state the emotion of the speaker even in challenging environments. Globally, there are two feature extraction approaches. The first one consists of using or finding single features of high performance [13,14]. The second approach is combinatorial in nature and makes use of different features to form large features sets. This is known as early or feature-level fusion. Table 1 provides a non-exhaustive list of works related to this topic. In terms of speech emotional feature extraction, various features that are based on a feature-level fusion have been investigated. Among these works, authors in [15] proposed a feature set that combines, among others, Harmonics-to-Noise Ratio (HNR), formants, Mel Frequency Cepstral Coefficients (MFCCs), and 19 channel filter bank (VOC19) features. In [16], acoustic features were combined with lexical features extracted from word transcripts obtained using an Automatic Speech Recognition (ASR) system: Bag-Of-Words (BOW). They achieved a four emotion recognition accuracy of 65.7% using Support Vector Machines (SVM). In [17], the authors extracted 286 features from speech signals that were exposed to babble noise with different signal-to-noise ratio levels, including formants, pitch, energy, MFCCs, Perceptual Linear Prediction (PLP), and Linear Predictive Coding (LPC), and were used for emotion classification by means of Naive Bayes (NB), K-Nearest Neighbors (KNN), the Gaussian Mixture Model (GMM), Artificial Neural Network (ANN), and SVM classifiers. They showed that NB and SVM classifiers provided the best results. In [18], 988 statistical functionals and regression coefficients were extracted from eight kinds of low-level features that were: intensity, loudness, MFCC, Line Spectral Pairs (LSP), Zero Crossing Rate (ZCR), probability of voicing, fundamental frequency F0, and F0 envelope. The proposed set provided a good detection rate of 83.10% for Speaker Independent (SI) SER on the Berlin Database. In [19], sub-band spectral centroid Weighted Wavelet Packet Cepstral Coefficients (W-WPCC) were proposed and were fused with Wavelet Packet Cepstral Coefficients (WPCC) and prosodic and voice quality features to deal with white Gaussian noise. In [20], Linear Predictive Cepstral Coefficients (LPCC) and MFCCs were derived from wavelet sub-bands and were fused with baseline LPCCs and MFCCs. The resulting feature dimension was reduced using the vector quantization method, and the obtained feature vector was used as input to a Radial Basis Function Neural Network (RBFNN) classifier. Recently, in [21], the authors proposed a combination of Empirical Mode Decomposition (EMD) with the Teager–Kaiser Energy Operator (TKEO). They proposed novel features named Modulation Spectral (MS) features and Modulation Frequency Features (MFF) based on the AM-FMmodulation model and combined them with cepstral features.

Table 1. Overview of Speech Emotion Recognition (SER) works based on feature-level fusion. HNR, Harmonics-to-Noise Ratio; EMO-DB, Berlin German Emotional Database; ZCR, Zero Crossing Rate; SI, Speaker Independent; IEMOCAP, Interactive Emotional Dyadic Motion Capture; PLP, Perceptual Linear Prediction; LPC, Linear Predictive Coding; LSP, Line Spectral Pairs; W-WPCC, Weighted Wavelet Packet Cepstral Coefficients; EMD-TKEO, Empirical Mode Decomposition with the Teager–Kaiser Energy Operator.

Work	Features	Classifier	Database	Speaker Dependency
[15]	Energy, pitch, jitter, shimmer, spectral flux, spectral centroid, spectral roll-off, intonation, intensity, formants, HNR, MFCC, VOC19	SVM	DES EMO-DB SUSAS	SD
[22]	MFCC, LPCC, ZCR, spectral roll-off, spectral centroid	SVM	EMO-DB	SI
[23]	Energy, pitch, formants, ZCR, harmony	NB	EMO-DB	SI
[16]	Energy, pitch, formants, jitter, shimmer, MFCC, word-stem presence indicators, BOW sentiment categories	SVM	IEMOCAP	SI
[17]	Formants, pitch, energy, spectral, MFCC, PLP, LPC	NB KNN GMM ANN SVM	EMO-DB SES	SD
[24]	Pitch, energy, ZCR, LPC, MFCC	SVM	EMO-DB Japan Thai	SD
[25]	Rhythm, temporal	ANN	EMO-DB	SD
[18]	Intense, loudness, MFCC, LSP, ZCR, probability of voicing, F0, F0 envelope	SVM	EMO-DB	SI
[19]	F0, power, formants, W-WPCC, WPCC	SVM	EMO-DB	SD
[20]	LPCC, WLPCC, MFCC, WMFCC	RBFNN	EMO-DB SAVEE	SD
[21]	MS, MFF, EMD-TKEO	SVM RNN	EMO-DB Spanish	SD

3. Methodology

As introduced earlier, the idea of the present work is to investigate a feature driven approach to SER. The considered system is composed of three main stages, which are: feature extraction, dimensionality reduction, and classification.

3.1. Feature Extraction

Generally, the most commonly used features in SER are divided into three categories, namely: vocal tract, prosodic and excitation source features. Vocal tract features are obtained by analyzing the characteristics of the vocal tract, which is well reflected in the frequency domain. Prosodic features represent the overall quality of the speech and are extracted from longer speech segments like syllables, words, and sentences. Excitation source features are those used to represent glottal activity, mainly the vibration of vocal folds.

In this paper, we tried to find a set of features that have physical meaning. Since voice is produced and perceived by a human being, we made a combination of all families of features that model the mechanisms of production and perception of speech. More precisely, the considered feature set was chosen with three different feature families, which were pitch, related to vocal folds, MFCCs belonging to the perceptual speech family, and a multi-resolution based feature describing the spectral content of speech. MFCC was used as it is based on the human speech perception mechanism and employs a bank of Mel spaced triangular filters to model the human auditory system that perceives sound in a nonlinear frequency binning. The speech production system includes vocal cords and a vocal tract. The vocal cords generate sound waves by vibrating against each other, and the vocal tract modulates the resulting sound. However, the vocal tract is influenced by several factors such as articulation and emotions and shows greater variability. Hence, only features that are related to vocal cords were used. All of the considered features are described in the following.

3.1.1. Pitch

Pitch, also known as fundamental frequency F0, is a feature that corresponds to the frequency of vibration of the vocal folds. There are many algorithms for computing the fundamental frequency of speech signals, and they are generally referred to as pitch detection algorithms. They can operate either in time or frequency domain. Frequency domain pitch estimators usually utilize Fast Fourier Transform (FFT) to convert the signal to the frequency spectrum. Time domain approaches are typically less computationally expensive. Major pitch estimation algorithms can be decomposed into two major steps: the first one finds potential F0 candidates for each window, and the second one selects the best ones. In this work, pitch was estimated by using the Robust Algorithm for Pitch Tracking (RAPT) [26]. The algorithm first computes the Normalized Cross-Correlation Function (NCCF) of a low-sample signal and records the locations of the local maxima in this first pass NCCF. The NCCF [27] is defined for a K sample length at lag k and analysis frame $x(n)$, $0 \leq n \leq N-1$, as:

$$NCCF(k) = \frac{\sum_{n=0}^{N-k} x(n)x(n+k)}{\sqrt{e_0 e_k}}, \tag{1}$$

having

$$e_k = \sum_{n=k}^{k+N-K} x(n)^2, k = 0, \ldots, K-1, \tag{2}$$

where N is the number of frames and k is the lag number. Next, NCFF is performed on the high-sample rate signal in the vicinity of the peaks found in the first pass. This generates a list of several F0 candidates for the input frame. Finally, dynamic programming is used for the final pitch estimate selection.

3.1.2. MFCC

MFCC [28] is an audio feature extraction that is extensively used in many speech related tasks. It is known that it was developed to mimic the human auditory perception. The MFCC feature extraction process is given as follows:

1. Pre-emphasize the speech signal: The pre-emphasis filter is a special kind of Finite Impulse Response (FIR), and its transfer function is described as:

   $$H(z) = 1 - \alpha z^{-1}, \tag{3}$$

   where α is the parameter that controls the slope of the filter and is usually chosen between 0.4 and 1 [29]. In this paper, its value was set to 0.97.
2. Divide the speech signal into a sequence of frames that are N samples long. An overlap between the frames was allowed to avoid the difference between the frames. Windowing was then is applied over each of the frames to reduce the spectral leakage effect at the beginning and end of each frame. Here, the Hamming window was applied, which was defined as $w(n) = 0.54 - 0.46\cos(2\pi n/N - 1)$.
3. Compute the magnitude spectrum for each windowed frame by applying FFT.
4. The Mel spectrum was computed by passing the resulting frequency spectrum through a Mel filter bank with a triangular bandpass frequency response.

5. Discrete Cosine Transform (DCT) was applied to the log Mel spectrum to derive the desired MFCCs.

3.1.3. Wavelet Based Feature Extraction Method

Discrete Wavelet Transform (DWT) has been adopted for a huge variety of applications, from speaker recognition [4] to Parkinson's disease detection [30]. Briefly, DWT is a time-scale representation technique that iteratively transforms the input signal into multi-resolution subsets of coefficients through high-pass and low-pass filters and decimation operators. Practically, DWT is performed by an algorithm known as sub-band coding or Mallat's algorithm. According to [31], a discrete signal $x(n)$ can be decomposed as:

$$x(n) = \sum_k a_{j_0,k}\phi_{j_0,k}(n) + \sum_{j=j_0} \sum_k d_{j,k}\psi_{j,k}(n), \tag{4}$$

where $\phi_{j_0,k}(n) = 2^{j_0/2}\phi(2^{j_0}n - k)$ is the scaling function at a scale of 2^{j_0} shifted by k, $\psi_{j,k}(n) = 2^{j/2}\psi(2^j n - k)$ is the mother wavelet at a scale of 2^j shifted by k, $a_{j_0,k}$ is the approximation coefficients at a scale of 2^{j_0}, and $d_{j,k}$ is the detail coefficients at a scale of 2^j. In this paper, MFCCs were extracted from DWT sub-band coefficients to produce DMFCC features. Figure 1 summarizes the DMFCC feature extraction process.

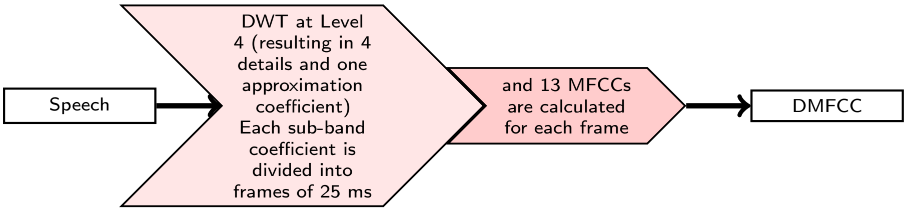
Speech
DWT at Level 4 (resulting in 4 details and one approximation coefficient) Each sub-band coefficient is divided into frames of 25 ms
and 13 MFCCs are calculated for each frame
DMFCC

Figure 1. Wavelet based feature extraction method.

After feature extraction, the sequence of features of each utterance was mapped into a global descriptor representative of the entire utterance. The most used mapping approach in emotion analysis is the Statistics Based Mapping Algorithm (SBMA). It consists of computing several statistics over the entire sequence. The resulting feature vector is used for analysis, learning, and classification. The computed statistics from each of the considered features are shown in Figure 2. For DMFCC, the statistics were computed over each of the MFCC matrices that were extracted from the wavelet sub-band coefficients.

The normalization has an important role in classification methods. Feature normalization guarantees that all the features will have the same scale [32]. Moreover, it is used to remove the speaker variability while preserving the discrimination between emotional classes. In this context, all features were subject to z-score normalization hereafter.

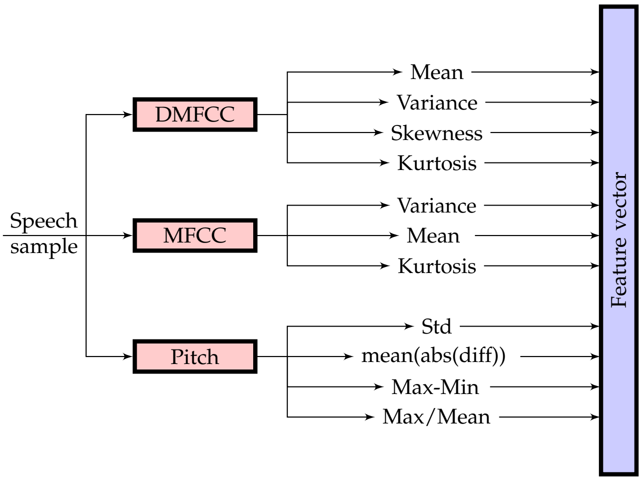

Figure 2. The used Statistics Based Mapping Algorithm (SBMA). Std stands for the standard deviation, and mean(abs(diff)) is the mean of the absolute value of the pitch's derivative. DMFCC, MFCCs derived from Discrete Wavelet Transform.

3.2. Dimensionality Reduction Using LDA

There are many techniques for reducing the dimensionality of the extracted feature vector with the goal of preserving the discriminability of the different emotion categories in the reduced dimensionality data. The most established techniques are Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA) [33]. Both of them reduce the dimensionality of features by projecting the original feature vector into a new subspace through a transformation. PCA optimizes the transformation by finding the largest variations in the original feature space, while LDA pursues the largest ratio of between-class variation and within-class variation when projecting the original feature to a subspace. In this paper, the dimensionality of the obtained feature vector through SBMA was further reduced using LDA. The main objective of LDA is to find a projection matrix W_{lda} that maximizes the so-called Fisher criterion:

$$W_{lda} = \arg\max_{W} \frac{|W^T S_b W|}{|W^T S_w W|}, \tag{5}$$

where S_b and S_w are the scatter between-class and within-class matrices, respectively, each defined as:

$$S_b = \frac{1}{g}\sum_{i=1}^{g}(\mu_i - \mu)(\mu_i - \mu)^T, \tag{6}$$

$$Sw = \frac{1}{g}\sum_{i=1}^{g}\frac{1}{\beta_i}\sum_{j=1}^{\beta_i}(v_{ij} - \mu_i)(v_{ij} - \mu_i)^T, \tag{7}$$

where g is the number of classes and μ_i and μ are the class mean and overall mean, respectively. v_{ij} are samples from class C_i, and β_i is the number of samples in class C_i.

3.3. Classification

NB and SVM are the typical generative and discriminative classification models, respectively. In this paper, we compare the two classifiers to examine their reliance in SER.

3.3.1. Naive Bayes Classifier

The NB classifier [34] is based on Bayes' theorem of conditional probability. The theorem states that the conditional probability that an event belongs to a class can be calculated from the conditional probabilities of finding particular events in each class and the unconditional probability of the event in each class. In other words, let X be the input data and H be some hypothesis that X belongs to class C. The conditional probability that an event belongs to a class C can be calculated by using the following equation:

$$P(H|X) = \frac{p(X|H)p(H)}{p(X)} \tag{8}$$

where $p(H)$ and $p(X)$ are the prior probabilities of H and X, respectively, and $P(X|H)$ is the posterior probability of X conditioned on H. $P(X|H)$ can be estimated as:

$$P(X|H) = \prod_{x \in X} p(x|H) \tag{9}$$

From this, the NB classifier is defined as:

$$C_{NB} = \underset{c \in H}{\text{argmax}}\, P(c) \prod_{x \in X} P(x|H) \tag{10}$$

3.3.2. SVM Classifier

SVM [35] is a binary classifier that separates input data into classes by fitting an optimal separating hyperplane to the training data in the feature space. It is extended to multi-class problems by using two strategies, which are: One-Versus-One (OVO) and One-Versus-All (OVA). In the OVA approach, each class is separated from the remaining ones. Thus, the number of SVMs that are trained equals the number of classes, and the final classification is determined by the highest score. The OVO approach, known also as pairwise classification, pairs each of the classes and trains an SVM for each pair. Each binary classifier is trained on only two classes; thus, the method constructs $g(g-1)/2$ binary classifiers, where g is the number of classes. For each test sample, the confidence level of each class is the majority voting result of these binary classifiers. The class with the most votes is selected as the final prediction. In our work, the OVO strategy was adopted. To obtain optimal performance of the SVM classifier, selection of a proper kernel function is essential. In this work, the linear kernel was exploited, since it does not need parameterization.

4. Results

We conducted SER experiments to evaluate the proposed approach, which was implemented using the algorithms described in Section 3. The evaluation was focused on its performance with and without noise incorporation. For bilingual emotion recognition, two databases were used in the experiments, one in English and the other in German.

4.1. Experimental Data and Parameters

In SER, the used datasets are categorized into three types. These are acted, authentic, and elicited emotional corpora [36].

- Acted: where the emotional speech is acted by subjects in a professional manner. The actor is asked to read a transcript with a predefined emotion.
- Authentic: where emotional speech is collected from recording conversations in real-life situations. Such situations include customer service calls and audio from video recordings in public places or from TV programs.
- Elicited: where the emotional speech is collected in an implicit way, in which the emotion is the natural reaction to a film or a guided conversation. The emotions are provoked, and experts label the utterances. The elicited speech is neither authentic nor acted.

In this paper, we aimed to investigate the performance of the proposed emotion recognition system with two types of corpora; IEMOCAP and EMO-DB, which are elicited and acted datasets, respectively.

4.1.1. IEMOCAP

IEMOCAP [37] is a multi-speaker and multimodal database, collected at the Speech Analysis and Interpretation Laboratory (SAIL) of the University of Southern California. It contains approximately twelve hours of audio-visual data from ten actors (five males and five females) and was recorded in five sessions. Each session had one male and one female performing improvisations or scripted scenarios designed to elicit particular emotions. The database was annotated by three annotators into several categorical labels, such as anger, happiness, excited, and so on. Only utterances with at least two agreed emotion labels were used for our experiments. Specifically, the categorical tags that we considered in the present work are: anger, excited, neutral, happiness, sadness, fear, and surprise. We merged happiness and excited as happiness, making the final dataset contain 7214 utterances (1031 anger, 1585 happiness, 1684 neutral, 1018 sadness, 33 fear, 90 surprise, and 1773 frustration).

4.1.2. EMO-DB

The German Emotional Speech Database (EMO-DB) [38] includes seven emotional states: anger, boredom, disgust, fear, happiness, and sadness, in addition to the neutral state. The utterances were produced by ten professional German actors (five females and five males) uttering ten sentences with an emotionally neutral content, but expressed with the seven different emotions. The total number of utterances was 535 divided among the seven emotional states: 128 anger, 83 boredom, 48 disgust, 71 fear, 72 happiness, 81 neutral, and 63 sadness.

4.2. Speaker Independent Experimental Results

Here, we present a series of experiments for SI analysis to understand the practical utility of the SER system in a real-world scenario. SI means that the speaker of the classified utterances was not included in the training database. In this context, we used Leave-One-Subject-Out (LOSO) cross-validation. That is, for both databases, the system was trained ten times, each time leaving one speaker out of the training set and testing the performance on the speaker left out.

SI systems offer several advantages. They are able to handle efficiently an unknown speaker. Thus, no training by the individual users was required. They also show a better generalization ability than the SDones, since they avoid overfitting. In addition, the experimental protocol in SI systems is deterministic, in the way that the exact configuration is known. In contrast to SD ones where cross-validation is employed, the random partitioning does not allow an exact reproduction of the configuration, making the results not directly comparable between research works.

The performance was evaluated in terms of accuracy, which is defined as:

$$Accuracy(\%) = \frac{Number\ of\ utterances\ that\ are\ correctly\ classified}{Total\ number\ of\ utterances} \times 100, \tag{11}$$

First of all, we conducted experiments in a clean environment. The motivation of the first experiment is twofold. First, it compared the performance of the SER system with individual features' families only and with fused features, respectively. Second, it compared the performance over the two used classifiers: NB and SVM. Figure 3 reports the obtained results.

The emotion recognition rates using EMO-DB for NB with pitch, MFCC, and DMFCC features were 46.23%, 64.79%, and 68.01%, respectively. Improvement was made by using the feature fusion technique, which gave the best recognition accuracy with a recognition rate of 82.32% when combining the three features. For the IEMOCAP database, the efficiencies of NB and SVM were 42.09% and 41.01%, respectively. By comparing these, we found that NB had the best recognition accuracy with the combination of MFCC and DMFCC. However, it still was nearly equal to that obtained when fusing the three considered features.

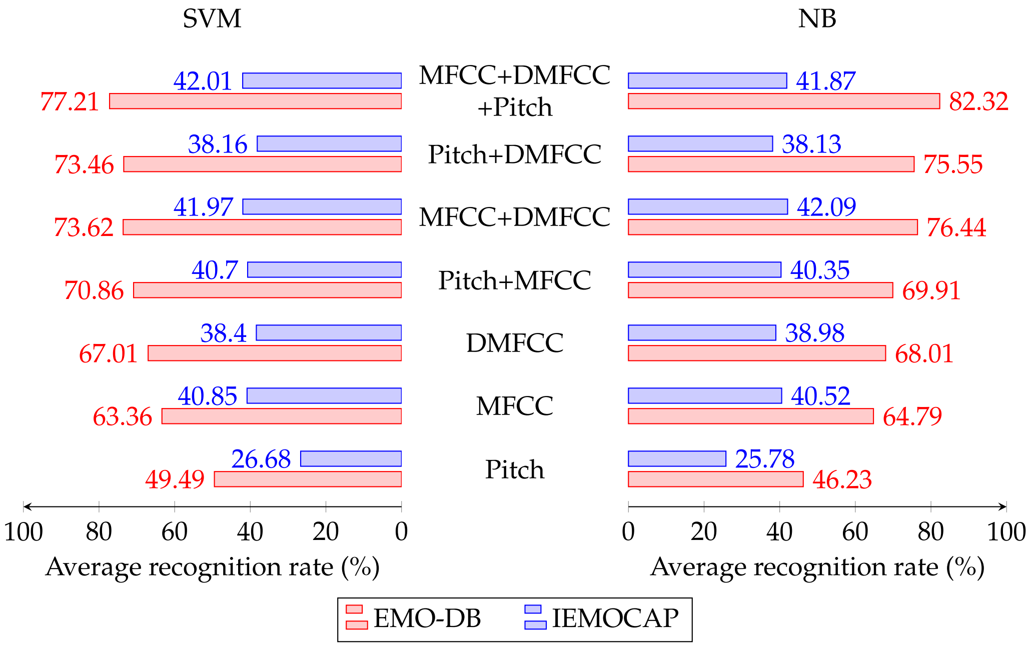

Figure 3. Obtained average accuracies for SI SER using the EMO-DB and IEMOCAP databases.

In order to further analyze the classification distribution of each emotion, the confusion matrices averaging the results of the ten separate experiments for SI SER experiments are shown in Tables 2 and 3. The values in the main diagonal give the average recognition accuracy of each emotion.

Table 2. Average confusion matrix of SI SER using EMO-DB (%).

	Anger	**Boredom**	**Disgust**	**Fear**	**Happiness**	**Neutral**	**Sadness**
Anger	92.10	0	0	1.60	5.69	0	0
Boredom	0	78.39	2.5	0.27	3	10.01	0
Disgust	0	10	74.25	1.25	3.75	3.25	0
Fear	0	0	3.98	80.10	12	0	0
Happiness	3.41	0	2.25	5.71	79.54	0	0
Neutral	0	19.03	0	0.91	0	74.22	1.43
Sadness	0	2.86	0	0	0	0	79.89

Table 3. Average confusion matrix of SI SER using IEMOCAP (%).

	Anger	**Happiness**	**Neutral**	**Sadness**	**Fear**	**Surprise**	**Frustration**
Anger	41	11.51	8.18	2.71	0	0.70	35.91
Happiness	9.07	30.52	19.23	9.05	0.12	0.93	31.07
Neutral	0.92	11.81	38.06	19	0.13	0.80	29.27
Sadness	0.45	2	19.06	63.73	0	1.28	13.49
Fear	7.5	29.64	9.93	0	0	0	22.93
Surprise	11.85	22.25	23.87	13.23	0	10.09	12.71
Frustration	9.88	11.03	20.77	8.22	0.05	0.72	49.33

As shown, using the IEMOCAP database, the emotions showed globally a high number of confusions. In EMO-DB, the emotions were relatively easily recognized, reaching the highest average accuracy of 92.10% for anger. This can be explained by the fact that EMO-DB is an acted database. Acted emotion expressions are generally more acoustically exaggerated than spontaneous ones [39], thus rendering them more easily differentiated. Moreover, taking an example of fear classification using IEMOCAP, a closer analysis revealed that fear was usually confused with frustration and happiness when annotators came to annotate the data. This behavior was in agreement with the one observed experimentally.

In real-world applications, SER nearly always involves capturing speech with background noise. The impact of such noise on speech signals and on SER performance may increase with the amount of traffic, the number of vehicles, their speed, and so on. The present behavioral experiment provides a performance analysis of the effects of such acoustic disturbances on SER. The considered background noises were based on the ones constructed for the AURORA noisy speech evaluation [40] and included: airport, train, babble, street, car, exhibition, and restaurant noise. Tables 4 and 5 provide the performance measurement in the presence noise for EMO-DB and IEMOCAP, respectively, at different Signal-to-Noise Ratio (SNR) levels.

Table 4. Average accuracies (%) of SI SER in the presence of different background noises (EMO-DB).

	NB							**SVM**						
	0 dB	**5 dB**	**10 dB**	**15 dB**	**20 dB**	**25 dB**	**30 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**	**20 dB**	**25 dB**	**30 dB**
Train	55.28	65.18	72.25	75.91	77.66	78.03	79.79	53.55	62.20	69.59	74.31	76.72	77.93	79.01
Exhibition	58.89	65.29	69.10	70.29	74.13	74.58	78.40	54.98	65.48	68.55	70.77	72.37	75.14	75.86
Street	54.60	66.51	72.25	76.09	79	79.65	81.73	52.60	62.72	69.69	72.98	75.56	78.31	78.90
Car	51.72	62.73	67.13	72.34	76.52	78.12	80.19	49.41	59.53	64.70	68.68	73.43	76.56	78.71
Restaurant	56.74	64.47	71.33	75.29	75.99	78.05	79.59	56.61	63.78	70.46	73.10	77.33	77.57	78.18
Babble	50.63	63.67	68.70	74.17	76.48	77.80	79.83	52.36	63.36	66.64	70.97	73.28	76.19	77.01
Airport	44.71	62.39	73.62	76.47	78.85	80.44	81.14	45.33	59.24	69.16	74.27	78.29	78.72	79.36
Average	53.22	64.32	70.63	74.36	76.95	78.10	80.10	52.12	62.33	68.40	72.15	75.28	77.20	78.15

Table 5. Average accuracies (%) of SI SER in the presence of different background noises (IEMOCAP).

	NB							**SVM**						
	0 dB	**5 dB**	**10 dB**	**15 dB**	**20 dB**	**25 dB**	**30 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**	**20 dB**	**25 dB**	**30 dB**
Train	36.98	37.99	39.18	40.17	40.91	41.36	41.69	37.22	38.45	39.77	40.43	40.97	41.31	41.84
Exhibition	33.60	34.78	35.98	37.12	37.73	38.69	39.69	33.95	34.72	36.00	36.77	38.08	38.84	39.66
Street	36.91	37.42	38.12	39.77	40.44	41.35	41.19	37.44	37.98	39.04	39.89	40.57	40.75	41.24
Car	37.02	37.21	38.10	38.85	39.42	39.84	40.65	37.15	37.18	37.99	38.78	39.43	39.90	40.65
Restaurant	34.95	35.68	36.68	37.22	38.33	39.52	40.42	36.13	36.63	37.17	37.42	38.68	39.86	40.64
Babble	34.71	35.98	37.19	38.58	39.63	40.34	41.03	36.06	36.93	37.97	38.76	39.56	40.24	41.03
Airport	37.00	38.00	38.97	39.86	40.96	41.61	41.56	37.41	38.54	39.44	40.16	40.76	41.18	41.77
Average	35.88	36.72	37.75	38.80	39.63	40.39	40.89	36.48	37.20	38.20	38.89	39.72	40.30	40.98

The reported results in Tables 4 and 5 for the three combined features showed again a greater performance of NB than SVM for EMO-DB. However, when using IEMOCAP, the performance of SVM was greater than that of NB from 0 dB to 10 dB and still nearly equal to that of NB for 15 dB and beyond. It was also noticed that NB was faster than SVM in terms of computational speed, which made it more suitable for real-world applications.

4.3. Comparison with Previous Related Work

As stated in Section 2, there is a number of works that have implemented an SER system based on a feature level fusion. However, research has been rarely devoted to SI SER, especially in the presence of noise. In [23], the authors performed a six class emotion recognition on EMO-DB. They used a feature set that combined pitch based features, energy, ZCR, duration, formants, and harmony features. The obtained recognition rates for each emotion were as follows: 52.7% for happiness with 33.9% misclassified as anger, 84.8% for boredom, 52.9% for neutral, 87.6% for sadness, 86.1% for anger, and 76.9% for fear. However, in our approach, we performed a seven class emotion recognition, reaching the recognition rates of 79.54% for happiness with only 3.41% misclassified as anger, 92.10% for anger, 78.39% for boredom, 74.22% for neutral, 79.89% for sadness, 92.10% for anger, and 80.10% for fear. In [22], all of the seven emotions of EMO-DB were used. The extracted feature set included MFCC, LPCC, ZCR, spectral roll-off, and spectral centroid, and an overall accuracy of 78.8% was achieved. More in particular, the recognition rates of each emotion were: 92.91% for anger, 74.68% for boredom, 68.42% for disgust, 70.91% for fear, 50% for happiness, 85.90% for neutral, and 90.57% for sadness. It seemed that sadness was easily recognized compared to our approach (79.89%), which was not the case of happiness, which we correctly recognized at an average rate of 79.54%. In [18], 988 statistical functionals and regression coefficients were extracted from eight kinds of low-level features. The proposed set provided an average recognition rate of 83.10% on EMO-DB. However, it was tested on a subset of only 494 speech samples out of 535. However, as demonstrated in Figure 3, we achieved a good accuracy using only 63 features when testing the performance on all of the 535 utterances of EMO-DB. Table 6 summarizes the obtained results for SI task on EMO-DB with that of literature works.

Table 6. Comparison with state-of-the-art works using EMO-DB.

Work	Anger	Happiness	Boredom	Neutral	Sadness	Fear	Disgust
[23]	86.1%	52.7%	84.8%	52.9%	87.6%	76.9%	-
[22]	92.91%	50%	74.68%	85.90%	90.57%	70.91%	68.42%
Our approach	92.10%	79.54%	78.39%	74.22%	79.89%	80.10%	74.25

4.4. Comparison with a Deep Learning Approach

Along with the advent of technology and the development of computer hardware, the applications of Deep Learning (DL) techniques has become a new research direction. In the field of SER, DL techniques have been recently proposed as an alternative to traditional ones. Many DL architectures have been introduced and are based on different models such as Recurrent Neural Networks (RNN) [41], Convolutional Neural Networks (CNN) [42,43], or Long Short Term Memory (LSTM) [44]. In this work, a CNN architecture was proposed for emotion recognition. In this study, feature extraction was applied first to input speech signals as described in Section 3 before they were input to the network. A conventional CNN is a multi-layer stacked neural network, which is built by stacking the following layers:

- Convolutional layer: utilizes a set of convolutional kernels (filters) to convert the input into feature maps.
- Non linearity: Between convolutional layers, an activation function is applied to the feature map to introduce nonlinearities into the network. Without this function, the network would essentially be a linear regression model and would struggle with complex data. In this paper, we used the most common activation function, which is the Rectified Linear Unit (ReLU) and defined as:

$$f(x) = max(0, x) \tag{12}$$

- Pooling layer: Its function is to decrease the feature maps size progressively to reduce the amount of parameters and computation in the network, hence to also control overfitting. Pooling involves selecting a pooling operation. Two common methods are average pooling and max pooling. In average pooling, the output is the average value of the feature map in a region determined by the kernel. Similarly, max pooling outputs the maximum value over a region in the feature maps. In this paper, the use of max pooling was adapted due to the property that the max operation preserves the largest activations in the feature maps.
- Softmax layer: Softmax regression is often implemented at the neural network's final layer for multi-class classification and gives the class probabilities pertaining to the different classes. Assuming that there are g classes, then its corresponding output probability for the jth class is:

$$Softmax(g_j) = \frac{\exp(g_j)}{\sum_{i=1}^{g} \exp(g_i)} \tag{13}$$

The overall architecture of the used CNN model is illustrated in Figure 4. The proposed model was built by stacking two convolutional layers, one fully connected layer and a softmax layer. The numbers of filters for the two convolutional layers were 64 and 128, respectively. Max pooling was added after convolution with a pooling length of two. The number of nodes in the fully connected layer was set to 128. A dropout layer was also added at the end of each layer to avoid overfitting. Details of CNN parameters are shown in Table 7.

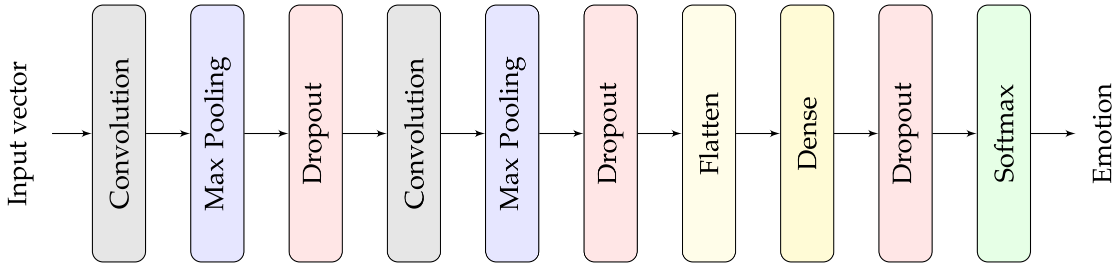

Figure 4. Block diagram of the overall architecture of the designed CNN model.

Table 7. Parameters of CNN.

Layer	Value
Convolution 1	64@12
Max Pooling 1	2
Convolution 2	128@12
Max Pooling 2	2
Dense layer	128
Dropout	0.1

The CNN parameters were optimized by RMSRprop with a learning rate of 0.00005. We used 100 epochs with a batch size of 16, which means the model saw the whole training data 100 times to update its weights.

Figure 5 illustrates the results of emotion classification using CNN and compares them to the results of conventional SVM and NB.

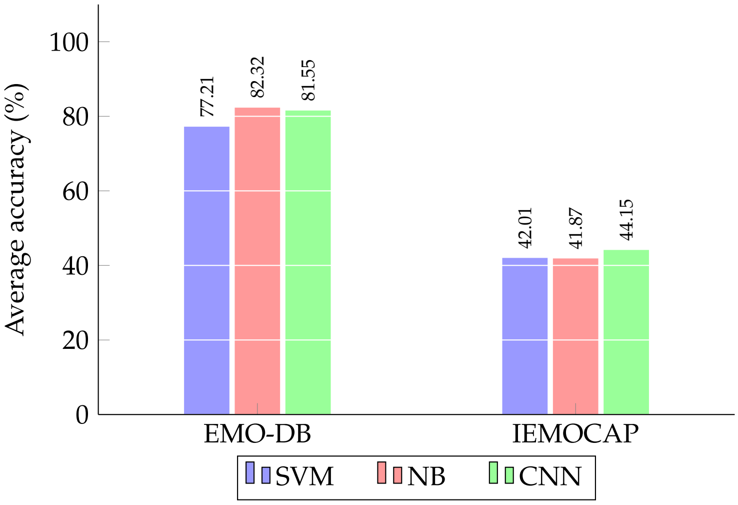

Figure 5. Comparison between ML and DL.

It can be seen that the CNN model achieved relatively ideal overall performance using IEMOCAP and improved the performance by 2.12%. However, the DL model was ineffective in being applied to EMO-DB and achieved an average accuracy of 81.55% compared to 82.32% for NB. These findings may be explained by the small size of EMO-DB. ML could still achieve a high classification accuracy in cases where the size of the available data was small, which demonstrated its powerful capacity for pattern recognition problems.

5. Conclusions

In this paper, the use of a feature fusion for SER was investigated in a noisy environment. Different features' combinations were attempted and showed improvements in classification compared to using individual features. Among the proposed combined features explored in this work, the combination of MFCC, MFCC derived from wavelets (DMFCC) and pitch based features remained the most reliable followed by the combination of MFCC and DMFCC. Comparison with state-of-the art works showed the possibility of using a relatively low-dimensional feature set with good accuracy. Since emotions do not have clear-cut boundaries (even people are usually confused at recognizing other people's emotions), there is a need to explore and develop classification methods that can handle this vague boundary problem. In this context, conventional ML algorithms were studied and compared with a DL technique. Experimental results demonstrated that when implementing DL for SER, one of the major challenges became a lack of availability of large datasets. To deal with the issue of limited datasets, one possibility for future research would be to consider data augmentation by either collecting or creating more data.

Author Contributions: Methodology, S.S.; Supervision, M.K., A.A. and S.B.J.; Writing—original draft, S.S.

Funding: This research received no external funding.

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

References

1. Al-Ali, A.K.H.; Dean, D.; Senadji, B.; Chandran, V.; Naik, G.R. Enhanced Forensic Speaker Verification Using a Combination of DWT and MFCC Feature Warping in the Presence of Noise and Reverberation Conditions. *IEEE Access* **2017**, *5*, 15400–15413. [CrossRef]
2. Al-Ali, A.K.H.; Senadji, B.; Naik, G.R. Enhanced forensic speaker verification using multi-run ICA in the presence of environmental noise and reverberation conditions. In Proceedings of the 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuching, Malaysia, 12–14 September 2017; pp. 174–179. [CrossRef]
3. Lee, M.; Lee, J.; Chang, J.H. Ensemble of jointly trained deep neural network based acoustic models for reverberant speech recognition. *Digit. Signal Process.* **2019**, *85*, 1–9. [CrossRef]
4. Sekkate, S.; Khalil, M.; Adib, A. Speaker Identification for OFDM-Based Aeronautical Communication System. *Circuits Syst. Signal Process.* **2019**, *38*, 3743–3761. [CrossRef]
5. Dhakal, P.; Damacharla, P.; Javaid, A.Y.; Devabhaktuni, V. A Near Real-Time Automatic Speaker Recognition Architecture for Voice-Based User Interface. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 504–520. [CrossRef]
6. Mallikarjunan, M.; Karmali Radha, P.; Bharath, K.P.; Muthu, R.K. Text-Independent Speaker Recognition in Clean and Noisy Backgrounds Using Modified VQ-LBG Algorithm. *Circuits Syst. Signal Process.* **2019**, *38*, 2810–2828. [CrossRef]
7. Xiaoqing, J.; Kewen, X.; Yongliang, L.; Jianchuan, B. Noisy speech emotion recognition using sample reconstruction and multiple-kernel learning. *J. China Univ. Posts Telecommun.* **2017**, *24*, 1–17. [CrossRef]
8. Staroniewicz, P.; Majewski, W. Polish Emotional Speech Database – Recording and Preliminary Validation. In *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions*; Esposito, A., Vích, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 42–49.
9. Tawari, A.; Trivedi, M.M. Speech Emotion Analysis in Noisy Real-World Environment. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 4605–4608. [CrossRef]
10. Huang, C.; Chen, G.; Yu, H.; Bao, Y.; Zhao, L. Speech Emotion Recognition under White Noise. *Arch. Acoust.* **2013**, *38*, 457–463. [CrossRef]
11. Hyun, K.; Kim, E.; Kwak, Y. Robust Speech Emotion Recognition Using Log Frequency Power Ratio. In Proceedings of the 2006 SICE-ICASE International Joint Conference, Busan, Korea, 18–21 October 2006; pp. 2586–2589. [CrossRef]
12. Yeh, L.Y.; Chi, T.S. Spectro-temporal modulations for robust speech emotion recognition. In Proceedings of the INTERSPEECH 2010, Makuhari, Japan, 26–30 September 2010.

13. Georgogiannis, A.; Digalakis, V. Speech Emotion Recognition using non-linear Teager energy based features in noisy environments. In Proceedings of the 2012 20th European Signal Processing Conference (EUSIPCO), Bucharest, Romania, 27–31 August 2012; pp. 2045–2049.
14. Bashirpour, M.; Geravanchizadeh, M. Speech Emotion Recognition Based on Power Normalized Cepstral Coefficients in Noisy Conditions. *Iran. J. Electr. Electron. Eng.* **2016**, *12*, 197–205.
15. Schuller, B.; Arsic, D.; Wallhoff, F.; Rigoll, G. Emotion Recognition in the Noise Applying Large Acoustic Feature Sets. In Proceedings of the Speech Prosody, Dresden, Germany, 2–5 May 2006.
16. Rozgic, V.; Ananthakrishnan, S.; Saleem, S.; Kumar, R.; Vembu, A.; Prasad, R. Emotion Recognition using Acoustic and Lexical Features. In Proceedings of the INTERSPEECH 2012, Portland, OR, USA, 9–13 September 2012; Volume 1.
17. Karimi, S.; Sedaaghi, M.H. Robust emotional speech classification in the presence of babble noise. *Int. J. Speech Technol.* **2013**, *16*, 215–227. [CrossRef]
18. Jin, Y.; Song, P.; Zheng, W.; Zhao, L. A feature selection and feature fusion combination method for speaker-independent speech emotion recognition. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 4808–4812. [CrossRef]
19. Huang, Y.; Tian, K.; Wu, A.; Zhang, G. Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition. *J. Ambient. Intell. Humaniz. Comput.* **2017**. [CrossRef]
20. Palo, H.K.; Mohanty, M.N. Wavelet based feature combination for recognition of emotions. *Ain Shams Eng. J.* **2018**, *9*, 1799–1806. [CrossRef]
21. Kerkeni, L.; Serrestou, Y.; Raoof, K.; Mbarki, M.; Mahjoub, M.A.; Cleder, C. Automatic Speech Emotion Recognition using an Optimal Combination of Features based on EMD-TKEO. *Speech Commun.* **2019**. [CrossRef]
22. Ruvolo, P.; Fasel, I.; Movellan, J.R. A learning approach to hierarchical feature selection and aggregation for audio classification. *Pattern Recognit. Lett.* **2010**, *31*, 1535–1542, doi:10.1016/j.patrec.2009.12.036. [CrossRef]
23. Yang, B.; Lugger, M. Emotion recognition from speech signals using new harmony features. *Signal Process.* **2010**, *90*, 1415–1423, doi:10.1016/j.sigpro.2009.09.009. [CrossRef]
24. Seehapoch, T.; Wongthanavasu, S. Speech emotion recognition using Support Vector Machines. In Proceedings of the 2013 5th International Conference on Knowledge and Smart Technology (KST), Chonburi, Thailand, 31 January–1 February 2013; pp. 86–91. [CrossRef]
25. Bhargava, M.; Polzehl, T. Improving Automatic Emotion Recognition from speech using Rhythm and Temporal feature. *arXiv* **2013**, arXiv:1303.1761.
26. Talkin, D. A robust algorithm for pitch tracking (RAPT). In *Speech Coding and Synthesis*; Klein, W.B., Palival, K.K., Eds.; Elsevier: Amsterdam, The Netherlands, 1995.
27. Kasi, K.; Zahorian, S.A. Yet Another Algorithm for Pitch Tracking. In Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA , 13–17 May 2002; Volume 1, pp. I-361–I-364. [CrossRef]
28. Davis, S.; MerMelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 357–366. [CrossRef]
29. Picone, J.W. Signal modeling techniques in speech recognition. *Proc. IEEE* **1993**, *81*, 1215–1247. [CrossRef]
30. Sakar, C.O.; Serbes, G.; Gunduz, A.; Tunc, H.C.; Nizam, H.; Sakar, B.E.; Tutuncu, M.; Aydin, T.; Isenkul, M.E.; Apaydin, H. A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. *Appl. Soft Comput.* **2019**, *74*, 255–263. [CrossRef]
31. Mallat, S. *A Wavelet Tour of Signal Processing*, 2nd ed.; Academic Press: San Diego, CA, USA, 1998.
32. Chul Min Lee.; Narayanan, S.S. Toward detecting emotions in spoken dialogs. *IEEE Trans. Speech Audio Process.* **2005**, *13*, 293–303. [CrossRef]
33. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179–188. [CrossRef]
34. Duda, R.; Hart, P. *Pattern Classifications and Scene Analysis*; John Wiley & Sons: New York, NY, USA, 1973.
35. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.:1022627411411. [CrossRef]
36. Ververidis, D.; Kotropoulos, C. Emotional speech recognition: Resources, features, and methods. *Speech Commun.* **2006**, *48*, 1162–1181. [CrossRef]

37. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335. [CrossRef]
38. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the INTERSPEECH ISCA, Lisbon, Portugal, 4–8 September 2005; pp. 1517–1520.
39. Zeng, Z.; Pantic, M.; Roisman, G.I.; Huang, T.S. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 39–58. [CrossRef] [PubMed]
40. Pearce, D.; Hirsch, H.G. The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions. In Proceedings of the ISCA ITRW ASR2000, Paris, France, 18–20 September 2000; pp. 29–32.
41. Tang, D.; Zeng, J.; Li, M. An End-to-End Deep Learning Framework for Speech Emotion Recognition of Atypical Individuals. In Proceedings of the INTERSPEECH 2018, Hyderabad, India, 2–6 September 2018; doi:10.21437/Interspeech.2018-2581. [CrossRef]
42. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **2019**, *47*, 312–323. [CrossRef]
43. Hossain, M.S.; Muhammad, G. Emotion recognition using deep learning approach from audio–visual emotional big data. *Inf. Fusion* **2019**, *49*, 69–78. [CrossRef]
44. Sarma, M.; Ghahremani, P.; Povey, D.; Goel, N.K.; Sarma, K.K.; Dehak, N. Emotion Identification from Raw Speech Signals Using DNNs. In Proceedings of the INTERSPEECH 2018, Hyderabad, India, 2–6 September 2018.