sensors

Communication

A Multimodal Approach to Improve Performance Evaluation of Call Center Agent

Abdelrahman Ahmed [1], Khaled Shaalan [2], Sergio Toral [1,*] and Yasser Hifny [3]

[1] Department of Electronic Engineering, University of Seville, 41092 Seville, Spain; abdahm@alum.us.es
[2] Faculty of Engineering & IT, British University in Dubai, Dubai 345015, United Arab Emirates; Khaled.shaalan@buid.ac.ae
[3] Faculty of Computers and Artificial Intelligence, University of Helwan, Helwan 11795, Egypt; yhifny@fci.helwan.edu.eg
* Correspondence: storal@us.es

Abstract: The paper proposes three modeling techniques to improve the performance evaluation of the call center agent. The first technique is speech processing supported by an attention layer for the agent's recorded calls. The speech comprises 65 features for the ultimate determination of the context of the call using the Open-Smile toolkit. The second technique uses the Max Weights Similarity (MWS) approach instead of the Softmax function in the attention layer to improve the classification accuracy. MWS function replaces the Softmax function for fine-tuning the output of the attention layer for processing text. It is formed by determining the similarity in the distance of input weights of the attention layer to the weights of the max vectors. The third technique combines the agent's recorded call speech with the corresponding transcribed text for binary classification. The speech modeling and text modeling are based on combinations of the Convolutional Neural Networks (CNNs) and Bi-directional Long-Short Term Memory (BiLSTMs). In this paper, the classification results for each model (text versus speech) are proposed and compared with the multimodal approach's results. The multimodal classification provided an improvement of (0.22%) compared with acoustic model and (1.7%) compared with text model.

Keywords: performance modeling; multimodal classification; BiLSTM; CNNs; attention layer

check for updates

Citation: Ahmed, A.; Shaalan, K.; Toral, S.; Hifny, Y. A Multimodal Approach to Improve Performance Evaluation of Call Center Agent. *Sensors* **2021**, *21*, 2720. https://doi.org/10.3390/s21082720

Received: 9 March 2021
Accepted: 9 April 2021
Published: 12 April 2021

1. Introduction

Evaluating the performance of call-center agents involves several issues. The first is that the evaluation is performed manually by listening to recorded calls and evaluating the content, which most likely will be a subjective evaluation [1,2]. Proficiency in oral communications is an essential skill in call centers, and it is very important for fulfilling the customers' needs. However, the customer service representative tone, oral proficiency, communications, and listening skills are most likely subjective factors that cause a bias in the evaluation process [3–5]. The second issue is that the number of calls is huge over a while, i.e., one year, which makes the manual evaluation very challenging. Hence, the evaluation is performed randomly over selected calls out of thousands of records. This can lead to missing the more realistic performance that occurred during the majority of the calls. The third obstacle is the diversity of the evaluators so that they may rank the same agent's performance differently. The lack of a unified system of evaluation can have a significant adverse impact on the business of call centers when the baseline is overlooked. Avoiding subjectivity and automating performance evaluation is essential in reducing the time and effort associated with the manual evaluation process. It leads to the establishment of a call center's performance baseline with a unified system of evaluation.

Objective methods in performance evaluation have been developed to overcome the subjective factors and assessment bias [6,7]. A discussion is presented for the two studies that considered the binary classification for either the text or speech of the recorded calls [8].

The productivity classification differs significantly when using speech processing instead of the text approach. There is a massive number of speech features that can be extracted from the recorded calls in comparison to the features of the text [9]. Furthermore, the text-based approach requires a minimum word error rate (WER) for the transcription system for better classification accuracy. However, more sophisticated data extraction and computational resources for speech modeling are required than for the text approach.

The research framework of this study is a multimodal classification based on different approaches that combine text and speech processing for improving accuracy. The proposed multimodal approach is the main paper contribution to empower the classification accuracy when combining the best classification performance obtained for speech and text. The models comprise different neural network structures to classify the speech utterances side by side with the corresponding call transcribed text into productive and nonproductive classes (binary classification). The study attempts to determine the best accuracy by combining the three techniques using the attention layer, Max Weights Similarity (MWS), and the multimodal system to improve the performance evaluation. Also, we investigated the performance of MWS compared with the Softmax function, which may reflect on the accuracy of the classification and encourage future studies.

The rest of the paper is structured as follows: Section 2 discusses the works related to performance evaluation in call centers. The study framework is illustrated in Section 3. The experiment and results are stated in Section 4. Finally, Section 5 concludes the paper and suggests the future research avenues.

2. Related Work

Many studies of machine learning are concerned with the processing of speech to detect the eminent factors of customer behavior and causes of complaints [10]. Other studies were concerned with analytics to detect the service quality based on the content of the recorded calls from customers. That framework uses Hadoop Map Reduce for text distances using Cosine distance and n-gram supported by slang words [11]. Perera et al. [12] studied the automatic performance evaluation of call center agents. They determined various factors to improve the performance evaluations, such as the speech utterances, the tone level, and the emotional characteristics, then classified using the support vector machine (SVM). Sudarsan et al. examined several systems to evaluate the performance based on prohibited words, emotional recognition, and others [13]. Their framework was based on platforms like Google, Wit, and Sphinx for transcription. Ahmed et al. [7] transcribed the text based on lexicon free Recurrent Neural Networks (RNN) supported by Connectionist Temporal Classification (CTC) objective function [14]. They annotated the corpus into productive/nonproductive and modeled the text using one generative approach (Naive Bayes), and two discriminative approaches (logistical regression and linear support vector machine (LSVM)) [6,7]. The generative and discriminative approaches were modeled on a bag of words as text features.

The emotion recognition and speech enhancement studies have an important exposure for behavior determination, and classification improvement [15–18]. They intended to enhance the speech quality to measure human behavior's emotional aspects and gestures based on sophisticated deep neural network training. Performance evaluation extends these studies by determining the performance from a human conversation. This study focuses on the call scenarios considering performance as the core part of the call center processes. The call center gives two advantages to the study: mixing the natural conversation between two parties and the high regularity in the conversation path over a high volume of calls. The call follows a standard and predefined script like welcoming message, agent name, and services [19,20].

Both CNNs and LSTMs dominate the deep learning approaches, and they have provided outstanding improvements in various studies [16,21]. Named entity recognition (NER) is a cascaded CNNs-LSTMs approach to extract critical medical information from electronic medical records [22]. The convolutional layers extract the prominent features in a fast and restricted manner. The long-short-term memory layers (LSTMs) are intended to handle the long sequential streams of inputs. The attention layer uses the Softmax function followed by the weighted average vector (context vector) and forwarded to the classifier to improve the accuracy [23,24].

Ahmed et al. explored the productivity measurement using speech signal processing [25]. The study was performed using MFCC 13 features extracted and forwarded to different combinations of CNNs and LSTMs structures. The attention layer was applied to enhance the binary classification up to 84.27%, which means an improvement of around 1.57% over text classification approaches. Besides, the attention weights highlighted the para-linguistic features where the productivity measurement takes place. Yet, there were issues associated with using MFCC features for the binary classification problem. MFCC only provides information about the vocal track; it ignores prosodic information. Therefore, providing MFCCs to CNNs is very restrictive and limits the ability of the CNN to use discriminative features for better classification. Accordingly, in this study, extended speech features were used to overcome the previous studies' limitations and improve the classification accuracy by combining the text and the speech models. The next sections demonstrate how several modeling approaches are combined to improve the model's accuracy and compare the previous text and speech classification approaches.

3. The Proposed Framework

There are several alternatives for modeling speech and text. In this study, two main schemes are proposed for modeling text and speech, as shown in Figure 1. The first branch is for modeling speech using CNNs, cascaded CNNs-LSTMs, and an attention layer. Each branch of the speech modeling presents one or more of the deep learning combinations, i.e., CNNs, CNNs-attention, CNNs-LSTMs, and CNNs-LSTMs-Attention layers. The features extraction has been extended to 65 features using the Open-smile toolkit [26]. Open-smile is a comprehensive toolkit for the extraction of audio and music features, and it supports low-level audio descriptors, such as Mel-frequency cepstral coefficients (MFCC), fundamental frequency, formant frequencies, perceptual linear predictive cepstral coefficients, CHROMA, and CENS features, loudness, line spectral frequencies, and linear predictive coefficients. The frames are forwarded to the four branches of speech modeling to get the best accuracy compared with other speech models. The text is transcribed by using an automatic speech recognition system with Word Error Rate (WER) 12.03% [8]. The transcribed text has been revised and edited manually to avoid WER that affects the text classification accuracy. The extraction of features uses a word embedding layer of approximately 4k vocabulary size (4930 words). The text branch follows the same speech neural network structure to attain the best accuracy for the four branches of the text. The models with the best accuracies for speech and text are then merged (concatenated) at the last neural network layer for sigmoid binary classification.

3.1. CNNs and BiLSTMs

CNNs are widely used in signal processing, and speech recognition tasks [27]. The CNNs help scan the extracted features' frames to obtain the best classification accuracy through the filters. This study considers two main branches as shown in Figure 1: one for the text features and another for the speech features. Each main branch is in turn divided into four subbranches that follows a similar scheme: Two of them make use of 1D-CNNs layers with *tanh* activation functions followed by either a max-pooling layer or an attention layer, and the other two make use of a 1D-CNN-BiLSTMs also followed by either a max-pooling layer or an attention layer. Finally, a logit sigmoid output layer performs the binary classification into a productive or nonproductive call. The combinations of different models were used in this study to identify the best classification performance.

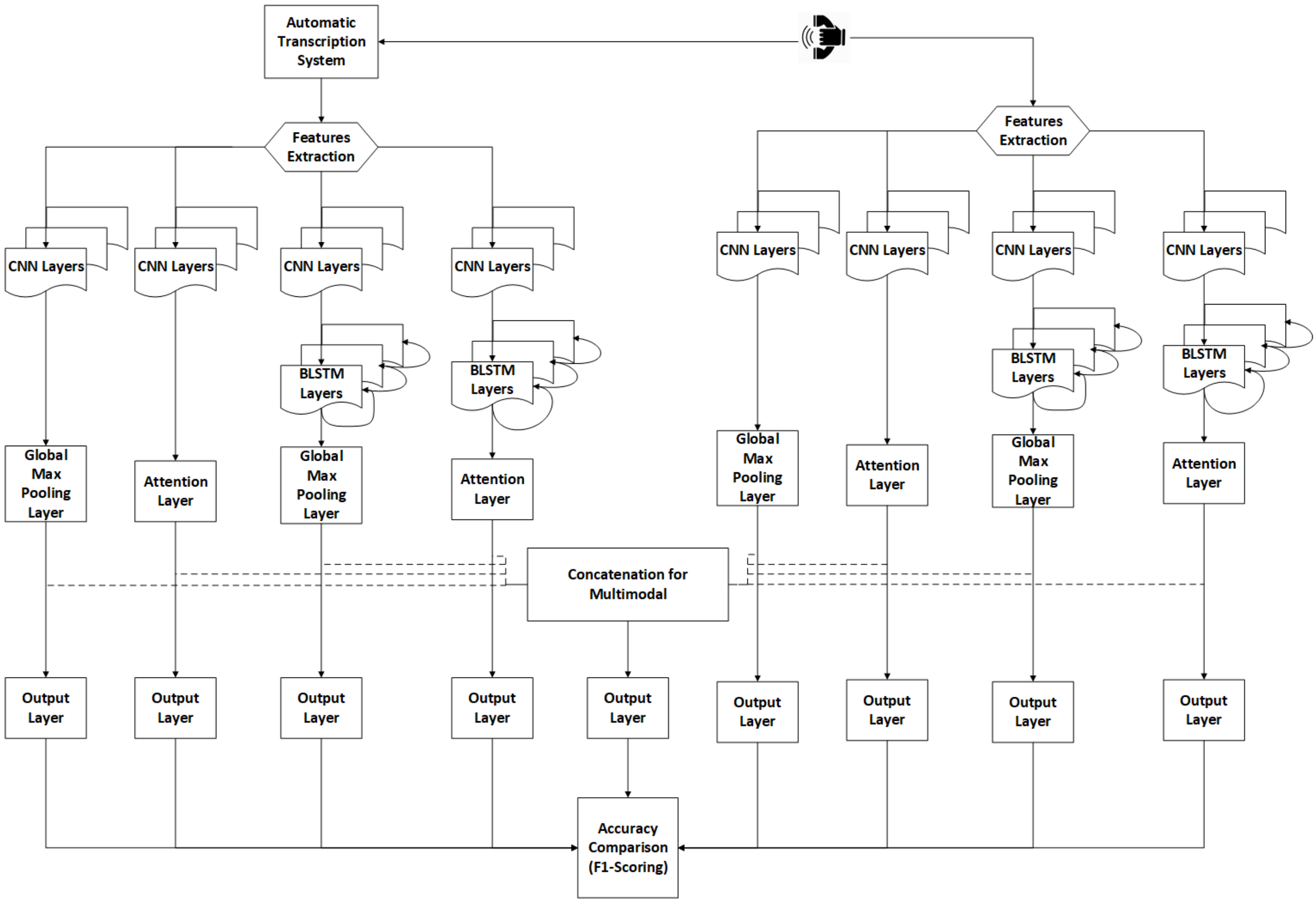

Figure 1. The proposed framework illustrates two schemes for speech and text. The dotted lines indicate the multimodal approach for merging one path for each scheme and forward it to the output layer.

3.2. Attention Layer

The sequence of vectors (frames) produced from CNN or LSTM and forwarded to the attention layer to convert them into a context vector [23,28,29]. The attention weight are forwarded to Softmax function at time t to generate the probability of the frame out of one to the remaining frames in the same speech segment. Then the context vector is generated by the weighted average of the frames probabilities. For each vector, $\mathbf{x}_t$ in a sequence of inputs, i.e., $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$, and the attention weights, α_t, are given by:

$$\alpha_t = \frac{\exp(f(\mathbf{x}_t))}{\sum_{j=1}^{T} \exp(f(\mathbf{x}_j))} \tag{1}$$

where $f(\mathbf{x}_t)$ is presented by the parameter $\mathbf{w}$ as follows:

$$f(\mathbf{x}_t) = \tanh(\mathbf{w}^T \mathbf{x}_t) \tag{2}$$

The weighted average of the Softmax generated weights and the input vector are summed to get the context vector C.

$$C = \sum_{t=1}^{T} \alpha_t \mathbf{x}_t \tag{3}$$

The Dense layer D uses $tanh$ activation function given by:

$$D = tanh(W^T C + b) \tag{4}$$

Being W are the hidden layers weights, and b is the bias. The Logit function is the output layer for two classes (productive/nonproductive).

$$y = Logit(D) \tag{5}$$

3.3. Max Weights Similarity (MWS)

The attention layer uses the Softmax function to determine the probability of the hidden layer weights among each other [23]. The Softmax function converts a vector of real values into probability values that sum up to one [30]. Sometimes, the Softmax function is referred to as multi-class logistic regression or the Softargmax function. For the speech processing branch, the wide variety in features (65 features × 25 ms frame, 10 ms frameshift) means that the Softmax can perform efficiently. However, in the text classification part, using a few embedded words limits its efficacy, so the classification accuracy is lower than in the speech processing branch. It can be explained because, in the text classification, the generated context vectors have values quite close to each other, so the attention layer does not have enough variability to reach a value that has a significant accuracy. The study proposes the Max Weights Similarity (MWS) function instead of the Softmax function to overcome the previous limitation. MWS aims to collapse the training weights around a reference value, which is the maximum value of the vector. The MWS function determines the similarity between the maximum value in the vector and the remaining values in the same vector. For each vector $\mathbf{x}_t$ in a sequence of inputs $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$, and $f(\mathbf{x}_t)$ in Equation (2), the attention weights α_t and maximum value β_t of the vector are given by:

$$\beta_t = \max(\exp(f(\mathbf{x}_1)), \exp(f(\mathbf{x}_2)), ..., \exp(f(\mathbf{x}_T))) \tag{6}$$

The cosine similarity equation for vectors a and b is as follows:

$$Cosine\_Similarity = \frac{a * b}{\|a\| * \|b\|} \tag{7}$$

The wights α_t of the context vector C in Equation (3) are given by:

$$\alpha_t = \frac{\exp(f(\mathbf{x}_t)) * \beta}{\|\exp(f(\mathbf{x}_t))\| * \|\beta\|} \tag{8}$$

Where $\|\exp(f(\mathbf{x}_t))\|$ is the normalized value of the vector of weights.

Then, the maximum value of the vector is chosen to give significant attention to the values of the vector compared with others. MWS will be applied either on the speech branch to compare its efficiency with the Softmax function.

3.4. Multimodal Approach

Many studies have been developed for deep-learning multimodal approaches [31–33]. In this study, merging the speech and text models are concatenated at the final layer for classification. The joint representation multimodal approach in [34] is applied to keep the speech and text features separated in the modeling process. Also, it gives a clear picture of the effect of using the multimodal compared with the same models trained alone. Figure 1 shows the dotted lines indicating the merging layer that combines the two speech and text models from the main branch. More specifically, different combinations of speech and text models are merged until achieving the best accuracy. Then, the five cross-validations and F1-scoring are applied for validation. In Equation (4), a merged dense layer concatenates the dense activation output from text and speech branches in Equation (9).

$$D_{Merged\_Dense} = Concatenate(D_{Speech}, D_{Text}) \tag{9}$$

Then forward the merged dense to the classifier in Equation (5). The Dense size is the total number of units for both speech and text layers.

4. The Experiment

The experiment is performed over three stages, i.e., speech processing, text processing, and multi-model classification (indicated by dotted box). Five-folds cross-validation with F1-scoring was used to validate the proposed experiments. The training was performed using Nvidia GPUs. The classification models were performed using Tensor-Flow backend and Keras APIs. Figure 2 summarizes the neural network parameters used in the proposed scheme of Figure 1.

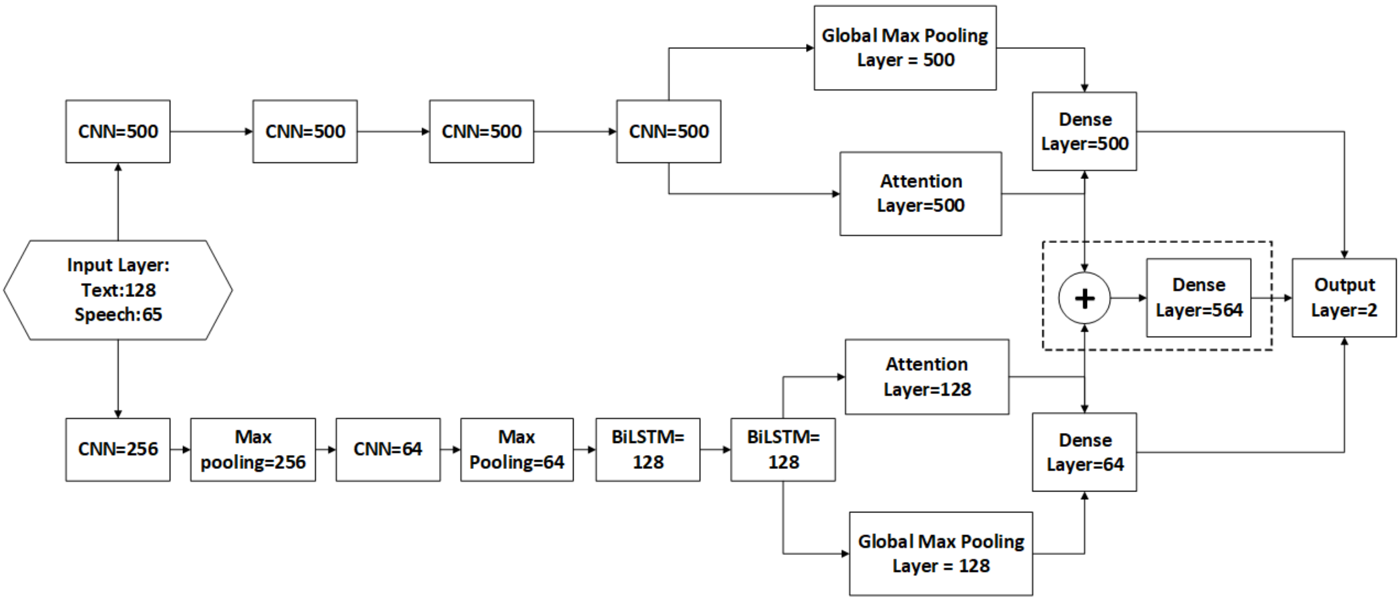

Figure 2. The Study Neural Networks Structure Units.

The neural network hyper-parameters are set following the configuration defined in the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [35], which is followed by various previous studies in emotional recognition and performance measurement [16,25].

4.1. The Data

Ethical approval has been granted for collecting the experiment corpus for research purposes from a real estate call center located in Egypt. A call recording system built-in VoIP call center was used between years 2014 and 2015 to collect real calls over landline phones with a sampling rate of 8 kHz. The selected random calls consist of seven hours over 30 calls (14 min per call on average), which is considered adequate compared to similar studies [16,25]. The corpus comprises six different agents between 25–35 years old; two females and four males. The calls were diarized , which is an algorithm to split the voice stream into smaller chunks. Speaker diarization is the process of splitting the speakers' utterances into separate segments [36] previously in [25] so that the talking time is 40% for females and 60% for males. The naming convention of the recorded calls is built from the metadata as Date, Time, Agent ID, Speaker ID (by the diariser), the call direction, Inbound, Outbound (The wave file name appears like DATE-TIME_AGENT-ID_SPK-ID_CALL-DIRECTION(INBOUND-OUTBOUND).wav). Three independent raters conducted a manual annotation process. The manual annotation may impact or bias the results because of the subjective performance evaluation of the raters. Hence, Krippendorff's Alpha is used to validate the agreement of the raters that should be more than 80%, which is achieved (Alpha > 0.79 in this study) [20].

4.2. Speech Processing

The Open-Smile toolkit for feature extraction [26] can be used to collect 65 features. It is based on INTERSPEECH 2016 Computational Paralinguistics Challenge (2016 COMPARE) [37]. It includes energy-related, spectral-related, and Low-Level Descriptors (LLDs); including logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness. The features are stated in Table 1. The resulting accuracy from training and validating the models are detailed in Table 2. The accuracy is compared with the previous study [25] in which 13 MFCC features were used.

Table 1. 65 provided Low-Level Descriptors (LLD).

54 spectral LLD
RASTA-style auditory spectrum MFCC 1–14 Spectral energy Spectral Roll Off Point Entropy, Spectral Flux, Skewness, Variance, Kurtosis, Slope, Harmonicity, Psychoacoustic Sharpness
7 voicing related LLD
Probability of voicing, F0 by SHS - Viterbi smoothing Jitter,logarithmic HNR, Shimmer PCM fftMag spectral Centroid SMA numeric
4 energy related LLD
Sum of auditory spectrum Sum of RASTA-style filtered auditory spectrum RMS Energy Zero-Crossing Rate

Table 2. Accuracy (Speech Processing) comparison.

Speech Accuracy % per Model Type		
Classification Method	**Type**	**Accuracy**
CNNs	MFCC	82.7%
CNNs-Attention	MFCC	84.27%
CNNs-BiLSTMs	MFCC	83.55%
CNNs-BiLSTMs-Attention	MFCC	83.54%
CNNs	LLD	90.1%
CNNs-Attention	LLD	92.48%
CNNs-Attention + MWS	LLD	92.88%
CNNs-BiLSTMs	LLD	92.67%
CNNs-BiLSTMs-Attention	LLD	92.68%
CNNs-BiLSTMs-Attention + MWS	LLD	92.25%

There was a significant improvement in speech classification using LLD than the previous study (MFCC), with the highest improvement being 8.4%. The attention layer supported with MWS gives a slight improvement of 0.2% for CNN compared with Softmax but about 0.18% less accuracy for CNN-LSTM.

4.3. Text Processing

The word embedding layer has been applied to indexed words for the transcribed Arabic text. The generated dictionary is around 4k words with a max stream length of 128 words. The same deep learning structure in Figure 1 was applied with the attention layer using Softmax and MWS. The results were compared with the results of previous experiments of text classification [7] using Logit and SVM based on a bag of words. The results are reported in Table 3.

Table 3. Accuracy (Text Processing) comparison.

Accuracy % per Model Type		
Classification Method	**Type**	**Accuracy**
Naive Bayes	Bag of words	67.3%
Logistic Regression	Bag of words	80.76%
Linear Support Vector Machine (LSVM)	Bag of words	82.69%
CNNs	Word Embedding	90.73%
CNNs-Attention	Word Embedding	90.98%
CNNs-Attention+MWS	Word Embedding	91.4%
CNNs-BiLSTMs	Word Embedding	89.87%
CNNs-BiLSTMs-Attention	Word Embedding	91.19%
CNNs-BiLSTMs-Attention+MWS	Word Embedding	91.12%

The deep learning text classification using the embedding of words shows a significant improvement of 8.7% over the generative and discriminative approaches using the bag of words. The MWS has higher accuracy than Softmax only for the CNNs approach (0.42%). (The same happened in the case of the speech approach). The CNNs-BiLSTM is less accurate than the CNNs-Attention model, which coincides with the results of a previous study [25]. It occurred because the BiLSTM is more efficient for long data streams, which is not the case in the short conversations in a call center. Accordingly, the attention layer does not provide a significant classification improvement in CNNs-BiLSTMs compared with CNNs.

4.4. Multimodal Approach (Speech + Text)

This step is required to increase the classification accuracy by combining (merging) the models at the final layer in Figure 1. The dotted box in Figure 2 is the merged dense $Merged\_Dense(Batch\_size, Param)$ of batch size = 32 with the following hyper-parameters:

$$Merged\_Dense(32, 564) = Conc(D_{Speech}(32, 500), D_{Text}(32, 64)) \quad (10)$$

The results reported in Table 4 and Figure 3.

Table 4. Accuracy (Multimodal models) comparison.

Multimodal Accuracy % per Model Type		
Text Model	**Speech Model**	**Accuracy**
CNNs	CNNs	90.44%
CNNs-Attention	CNN	90.1%
CNN	CNNs-Attention	92.63%
CNN	CNNs-Attention + MWS	92.9%
CNN-Attention	CNNs-Attention	91.76%
CNN-Attention + MWS	CNNs-Attention + MWS	93.1%
CNNs	CNNs-BiLSTMs-Attention	91.8%
CNNs	CNNs-BiLSTMs-Attention + MWS	91.9%
CNNs-Attention	CNNs-BiLSTMs	90.36%
CNNs-Attention + MWS	CNNs-BiLSTMs	91.1%
CNNs-Attention	CNNs-BiLSTMs-Attention	91%
CNNs-Attention + MWS	CNNs-BiLSTMs-Attention + MWS	91.1%

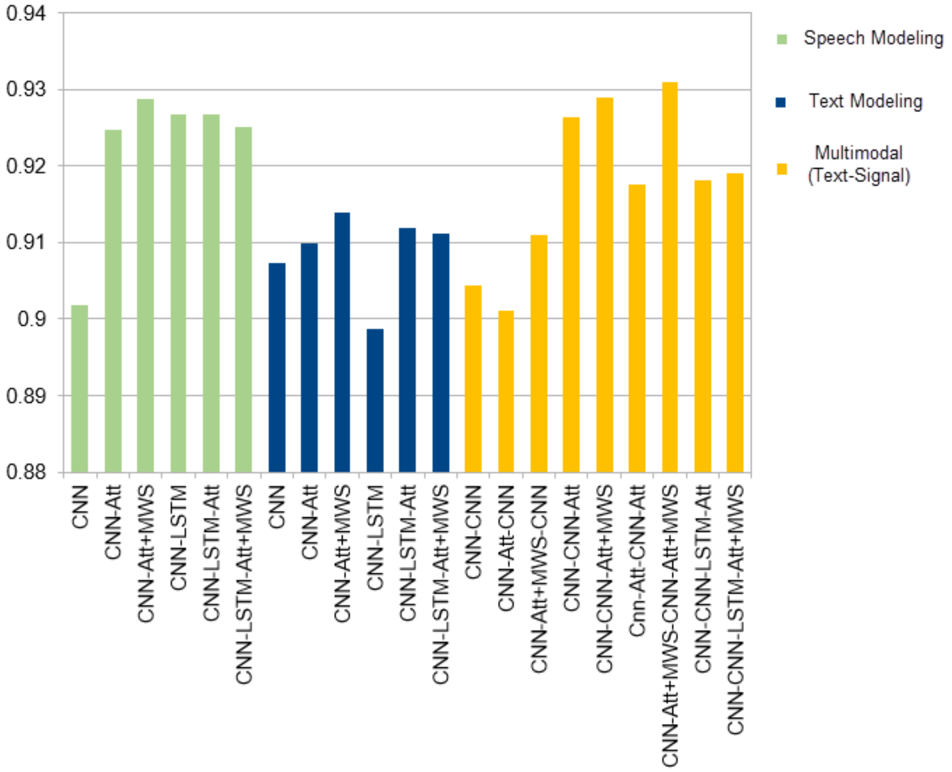

Figure 3. Modeling Approaches.

As shown in Figure 3, the multimodal approach provides a better classification accuracy by combining the CNNs-attention model for speech features and the CNNs-attention model for text features; both are implemented with the MWS function instead of the Softmax function. The Multimodal MWS approach had an improvement of 0.22% for modeling speech and 1.7% for modeling text. The accuracy of the multimodal classification for MWS was slightly better than that of Softmax by 1.34% for the same model. Findings reveal that the multimodal approach improves previous approaches and not combined models. However, we propose several lines in which this study could be extended for higher classification accuracy: (1) extending the vocabulary of the text model, in case of using the automatic transcription system, to reduce the Out Of Vocabulary (OOV) impact on productivity measurement, (2) improving the text model using pre-training approaches, i.e., Glove and BERT [38,39]. However, the previous pre-trained models do not support the Arabic language in the call centers domain, which requires more effort for data collection and training, (3) investigating other multimodal approaches like Coordinated-representation of structured space for merging the models [34].

Table 5 summarizes the results using MWS and Softmax functions for the models with the highest accuracy.

Table 5. The Table Compares the MWS method with Softmax used in Attention Layer.

MWS vs. Softmax—Accuracy Improvement%			
Method	**Speech Model**	**Text Model**	**Multimodal**
Softmax	92.68%	90.98%	91.76%
MWS	92.88%	91.4%	93.1%
Delta	0.2%	0.42%	1.34%

5. Conclusions

The automatization of the call center's performance measurement is a critical task due to subjective evaluation. A novel method is proposed based on the multimodal approach by merging the speech and text models. The experiment was conducted over seven hours of speech at the real estate call center. In the study, 65 features of speech were applied using the Open-smile feature extraction instead of MFCC, and a significant improvement of 8.4% was achieved. The deep learning approaches for learning text improved the accuracy by 8.7% compared with the generative and discriminative approaches. The final multimodal approach achieved is 93.1%, which was an approximate improvement of about 1.7% over text classification and about a 0.22% improvement over speech processing. The Max weights similarity (MWS) method gave a minor improvement compared with the Softmax function, which is recommended for further investigation over different domains. It is recommended that future researchers extend this study by using the Bert context extraction for modeling text. Besides, applying various multimodal approaches in the early stages is worth investigating to improve the classifications.

Author Contributions: Conceptualization, A.A., Y.H. and S.T.; methodology, A.A., S.T. and Y.H.; software, A.A. and Y.H.; validation, A.A. and Y.H.; writing—original draft preparation, A.A.; visualization, A.A.; supervision, S.T. and K.S.; project administration, A.A. and Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: There is no funders have contributed in any stage of the study.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: Special thanks to Luminous Technologies for the training corpus and modeling servers.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

1. Breuer, K.; Nieken, P.; Sliwka, D. Social ties and subjective performance evaluations: An empirical investigation. *Rev. Manag. Sci.* **2013**, *7*, 141–157. [CrossRef]
2. Dhanpat, N.; Modau, F.D.; Lugisani, P.; Mabojane, R.; Phiri, M. Exploring employee retention and intention to leave within a call center. *SA J. Hum. Resour. Manag.* **2018**, *16*, 1–13. [CrossRef]
3. Frederiksen, A.; Lange, F.; Kriechel, B. Subjective performance evaluations and employee careers. *J. Econ. Behav. Organ.* **2017**, *134*, 408–429. [CrossRef]
4. Gonzalez-Benito, O.; Gonzalez-Benito, J. Cultural vs. operational market orientation and objective vs. subjective performance: Perspective of production and operations. *Ind. Mark. Manag.* **2005**, *34*, 797–829. [CrossRef]
5. Echchakoui, S.; Baakil, D. Emotional Exhaustion in Offshore Call Centers: A Comparative Study. *J. Glob. Mark.* **2019**, *32*, 17–36. [CrossRef]
6. Ahmed, A.; Hifny, Y.; Toral, S.; Shaalan, K., A Call Center Agent Productivity Modeling Using Discriminative Approaches. In *Intelligent Natural Language Processing: Trends and Applications*; Book Section 1; Springer: Berlin/Heidelberg, Germany, 2018; pp. 501–520.
7. Ahmed, A.; Toral, S.; Shaalan, K. Agent productivity measurement in call center using machine learning. In *International Conference on Advanced Intelligent Systems and Informatics*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 160–169.
8. Ahmed, A.; Hifny, Y.; Shaalan, K.; Toral, S. End-to-End Lexicon Free Arabic Speech Recognition Using Recurrent Neural Networks. *Comput. Linguist. Speech Image Process. Arab. Lang.* **2018**, *4*, 231.
9. Dave, N. Feature extraction methods LPC, PLP and MFCC in speech recognition. *Int. J. Adv. Res. Eng. Technol.* **2013**, *1*, 1–4.
10. Bae, S.M.; Ha, S.H.; Park, S.C. A web-based system for analyzing the voices of call center customers in the service industry. *Expert Syst. Appl.* **2005**, *28*, 29–41. [CrossRef]
11. Karakus, B.; Aydin, G. Call center performance evaluation using big data analytics. In Proceedings of the 2016 International Symposium on Networks, Computers and Communications (ISNCC), Hammamet, Tunisia, 11–13 May 2016; pp. 1–6.
12. Perera, K.N.N.; Priyadarshana, Y.; Gunathunga, K.; Ranathunga, L.; Karunarathne, P.; Thanthriwatta, T. Automatic Evaluation Software for Contact Centre Agents' voice Handling Performance. *Int. J. Sci. Res. Publ.* **2019**, *5*, 1–8.
13. Sudarsan, V.; Kumar, G. Voice call analytics using natural language processing. *Int. J. Stat. Appl. Math.* **2019**, *4*, 133–136.
14. Ahmed, A.; Hifny, Y.; Shaalan, K.; Toral, S. Lexicon free Arabic speech recognition recipe. In *International Conference on Advanced Intelligent Systems and Informatics*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 147–159.
15. Neumann, M.; Vu, N.T. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. *arXiv* **2017**, arXiv:1706.00612.

16. Hifny, Y.; Ali, A. Efficient Arabic Emotion Recognition Using Deep Neural Networks. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6710–6714.
17. Cho, J.; Pappagari, R.; Kulkarni, P.; Villalba, J.; Carmiel, Y.; Dehak, N. Deep neural networks for emotion recognition combining audio and transcripts. *arXiv* **2019**, arXiv:1911.00432.
18. Li, P.; Jiang, Z.; Yin, S.; Song, D.; Ouyang, P.; Liu, L.; Wei, S. PAGAN: A Phase-Adapted Generative Adversarial Networks for Speech Enhancement. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–9 May 2020; pp. 6234–6238.
19. Cleveland, B. *Call Center Management on Fast Forward: Succeeding in the New Era of Customer Relationships*; ICMI Press:Colorado Springs, CO, USA, 2012.
20. Hayes, A.F.; Krippendorff, K. Answering the call for a standard reliability measure for coding data. *Commun. Methods Meas.* **2007**, *1*, 77–89. [CrossRef]
21. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204.
22. Li, L.; Xu, W.; Yu, H. Character-level neural network model based on Nadam optimization and its application in clinical concept extraction. *Neurocomputing* **2020**, *414*, 182–190. [CrossRef]
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Los Angeles, CA, USA, 8–9 December 2017.
24. Zhang, M.; Yang, Y.; Ji, Y.; Xie, N.; Shen, F. Recurrent attention network using spatial-temporal relations for action recognition. *Signal Process.* **2018**, *145*, 137–145. [CrossRef]
25. Ahmed, A.; Toral, S.; Shaalan, K.; Hifny, Y. Agent Productivity Modeling in a Call Center Domain Using Attentive Convolutional Neural Networks. *Sensors* **2020**, *20*, 5489. [CrossRef]
26. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia, Nice, France, 21–25 October 2010; pp. 1459–1462.
27. Palaz, D.; Collobert, R. *Analysis of CNN-Based Speech Recognition System Using Raw Speech as Input*; Technical Report; Idiap: Martigny, Switzerlnad, 2015.
28. Norouzian, A.; Mazoure, B.; Connolly, D.; Willett, D. Exploring attention mechanism for acoustic-based classification of speech utterances into system-directed and non-system-directed. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 7310–7314.
29. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y.N. Convolutional sequence to sequence learning. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, Sydney, Australia, 17 July 2017; pp. 1243–1252.
30. Bridle, J.S. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*; Springer: Berlin/Heidelberg, Germany, 1990; pp. 227–236.
31. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the ICML, Bellevue, WA, USA, 28 June–2 July 2011.
32. Ramachandram, D.; Taylor, G.W. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Process. Mag.* **2017**, *34*, 96–108. [CrossRef]
33. Kahou, S.E.; Bouthillier, X.; Lamblin, P.; Gulcehre, C.; Michalski, V.; Konda, K.; Jean, S.; Froumenty, P.; Dauphin, Y.; Boulanger-Lewandowski, N.; others. Emonets: Multimodal deep learning approaches for emotion recognition in video. *J. Multimodal User Interfaces* **2016**, *10*, 99–111. [CrossRef]
34. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [CrossRef]
35. 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*, 335–359. [CrossRef]
36. Broux, P.-A.; Desnous, F.; Larcher, A.; Petitrenaud, S.; Carrive, J.; Meignier, S. S4D: Speaker Diarization Toolkit in Python. *Interspeech* **2018**. [CrossRef]
37. Schuller, B.; Steidl, S.; Batliner, A.; Hirschberg, J.; Burgoon, J.K.; Baird, A.; Elkins, A.; Zhang, Y.; Coutinho, E.; Evanini, K.; et al. The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. In Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech 2016), San Francisco, CA, USA, 8–12 September 2016; Volume 1–5. pp. 2001–2005.
38. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
39. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.