MSMT-FN: Multi-segment Multitask Fusion Network for Marketing Audio Classification

Anonymous ICME submission

Abstract—Audio classification is an inherently multimodal task that is adopted in a diverse set of application scenarios, particularly in sentiment analysis or emotion recognition. In the field of marketing, classifying the attitude or propensity of potential customers in recorded marketing phone calls is an important first step toward business conversion. In this work, we propose a novel multi-segment multitask fusion network that is uniquely designed for addressing this business demand. Using both a proprietary dataset in phone marketing, and several open benchmark datasets in multimodal sentiment analysis, we show the effectiveness of our proposed fusion network with better performance when compared to the current state-of-the-art MMML and DF-ERC methods. Our codes and the MarketCalls dataset will be made public upon acceptance of the current paper.

Index Terms—Marketing Audio, Audio Classification, Multitask Learning, Multimodal Fusion

I. Introduction

Automatic voice-call robots have been increasingly adopted to replace human sales representatives to generate hundreds or even thousands of marketing phone calls each day for marketing purpose, immensely boosting marketing productivity. However, sales leads with high conversion potentials are invaluable, and existing manual classification approach incurs significant labor costs, making it hardly scalable. A solution to efficiently and effectively classify these large-volume phone call recordings is in dire needs.

To the best of our knowledge, few research has yet been conducted in classifying marketing phone call recordings, as such benchmark dataset has not been made publicly accessible. To enable future research in marketing audio classification, we curate a benchmark dataset containing around 1,000 marketing phone calls in Mandarin that are at least one minute in length, and call it MarketCalls. Unlike existing open benchmark datasets such as CMU-MOSI [1] and CMU-MOSEI [2], each sample in MarketCalls contains rounds of conversations between a sales representative and a target customer, and therefore rich contextual information to judge the propensity of the target customer. Also, the classification task is not utterance-based, which increases its challenges in classification. Finally, given that most existing audio datasets are in English, our curated dataset can enable the investigation of the generalizability in newly proposed audio classification methods trained on different languages.

We propose a novel and highly effective solution, namely Multi-segment Multitask Fusion Network (MSMT-FNet) in this work to classify these phone call recordings. In MSMT-FNet, we adopt data augmentation to enhance sample efficiency while reduce data annotation costs. Meanwhile, we consider acoustic signal as complementary to the textual signal in audio classification, and propose a network design using both cross-attention module and bottleneck fusion mechanism for modality fusion. Each audio sample is divided into multiple segments to better capture the fine-grained information within each segment. We apply a bi-directional GRU module to leverage the contextual information embedded among the segments of the same audio recording. Lastly, we put all of the above components under the multi-task learning framework to further enhance MSMT-FNet's generalizability and robustness in different tasks.

In order to evaluate the effectiveness of the MSMT-FNet, we conduct comprehensive experiments using both the MarketCalls dataset and three open benchmark datasets, including CMU-MOSI, CMU-MOSEI, and MELD [3]. Our contributions can be summarized as follows:

- We create a new benchmark dataset for audio classification in marketing and make it available for the research community. The audios were in mandarin, a non-English language that can enrich research opportunities in the audio domain.
- We propose a novel end-to-end classification network, namely MSMT-FNet, that is effective in leveraging the rich contextual signal embedded in lengthy dual conversations in marketing phone calls.
- We show that the MSMT-FNet achieves overall better performance in MarketCalls when compared to the replicated state-of-the-art baseline MMML method; and that MSMT-FNet can be generalized to multimodal sentiment analysis tasks using three additional open benchmark datasets.

II. Related Work

Most existing works related to audio classification tasks concentrated in multimodal sentiment analysis and emotion recognition [4]–[6]. Marketing audio classification also falls in these broad categories, although it infers a more specific emotional propensity toward the purchasing of a given service or product being offered in the marketing call by a salesperson. To our surprise, very few high quality works exist on this important topic, particularly in the field of marketing.

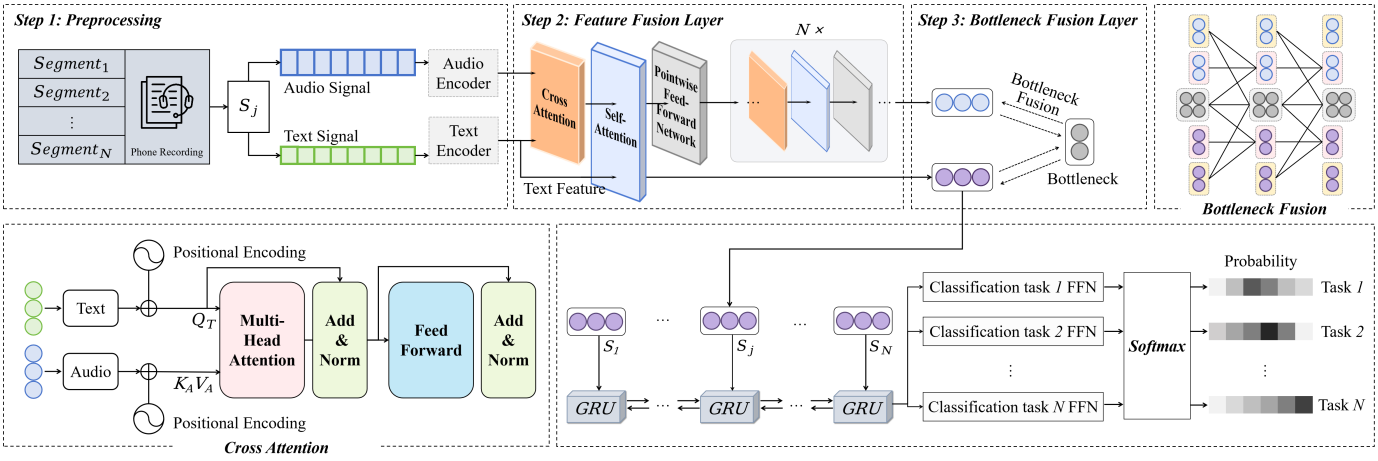

Fig. 1. MSMT-FNET Network Architecture. Step 1: Preprocessing. Each audio recording is broken into segments, and the audio and textual channels are extracted and encoded to obtain embeddings of the two channels. Step 2: Feature Fusion Layer. The textual channel serves as the backbone channel and is independently put through self-attention blocks; while a separate complementary channel is created by fusing the audio channel with the text channel using cross- and self-attention blocks. Step 3: Bottleneck Fusion Layer. A bottleneck fusion mechanism is adopted to more effectively fuse both channels from step 2. Finally, BiGRU is adopted under a multi-task learning framework for all segments within one audio recording to generate classification prediction for different tasks.

In contrast to more traditional approaches using audio signal feature extraction methods [7], most recent works applied attention-based approaches for both modality alignment [8], [9] and fusion [6], [10]–[13] in audio classification.

In alignment, Xu et al. proposed to use attention mechanism to learn the alignment between speech frames and text words to generate more accurate multimodal representations for emotion recognition from speech [8]. Goncalves et al. focused on addressing the challenges in emotion recognition given non-ideal conditions for audiovisual models, including misalignment of modalities, lack of temporal modeling, and missing features due to noise or occlusions. They combined auxiliary networks in a transformer architecture using an optimized training mechanism to obtain improved accuracy and robustness [9]. Yang et al. translated both the visual and audio features to textual features using BERT to tackle inferior feature qualities in visual and audio signals [11]. Similarly, Kim et al. proposed the All-modalities-in-One BERT to more effectively fuse multimodal signals for sentiment analysis [12].

In multimodal fusion, Zadeh et al. modeled the intra-modality and inter-modality dynamics using the proposed Tensor Fusion Network in an end-to-end approach [14]. Zheng et al. proposed to leverage transformer-based architecture to capture intra-modal and cross-modal interactions among multimodal signals [10]. Specifically, both single-modal and cross-modal transformers were adopted to unimodal and cross-modal features, and an audio-text-speaker fusion strategy was applied in fusion followed by attention mechanism to focus on important modalities. They also adopted a multi-head attention mechanism bidirectional GRU (MHA-GRU) to extract contextual information. Wu et al. proposed the multimodal multi-loss (MMML) fusion network [13] that consists of three major components: robust feature networks was applied to extract representations from each modality; cross-modal and self-attention mechanisms were combined to capture both intra- and inter-modality dependencies in fusion; and lastly a multi-loss training strategy was adopted to further enhance fusion effectiveness.

The above works have provided us with significant insights in our proposed approach in feature extraction and fusion. We also adopt self-attention and cross-modal attention mechanisms in feature learning and fusion. On top of these, we design a different fusion architecture with textual signal serving as the backbone signal channel, and combine it with bottleneck fusion strategy [15] for more effective learning. We also adopt bi-directional GRU to capture contexts in multi-segments within the same audio sample. Lastly, we employ a multi-task learning framework [16] to meet both unique business demand and enhance model robustness as well as its generalizability.

III. Method

A. Problem Formulation

Given a recorded audio containing conversations between a salesperson and a target customer for different products and services such as dental care and cosmetic products, we want to classify the customers into different categories according to their likelihood of conversion (e.g., very positive, neutral and receptive of the call, a little impatient and negative about the call but mostly remaining being polite, explicit refusal sometimes with

abusive language). These categories are manually labeled by experienced salespersons for model training purpose.

Let A_i denote the audio recording from the i-th target customer in our dataset. To better capture nuances of each exchange term, A_i can be split into conversation segments $a_j = \{a_{j,s}, a_{j,c}\}$ for $j \in \{1, 2, \ldots, l_i\}$, where l_i denotes the number of segments for A_i. Each segment a_j consists of a round of conversation between the salesperson s and the customer c. Let us denote the extracted representation from each audio recording A as $\mathbf{X} = \mathcal{F}(A)$, the goal of the current work is to obtain a model $\mathcal{N} : \mathbf{X} \rightarrow Y$, where Y denotes the purchasing propensity level.

B. Data Augmentation

To reduce costs in annotation and enhance sample efficiency as well as model robustness, we apply several augmentation strategies to the extracted acoustic and textual modalities to generate a much larger training set.

1) Audio Augmentation.: We apply the following methods to augment the audio signal in our dataset: 1) Gaussian background noise is added to the audio signal to simulate the real world environments with various sound levels [17]. 2) Speed perturbation is applied to modify the playback speed of the audio without altering its pitch [18]. 3) Random masking is adopted to mask or mute randomly selected portions of the audio signal [19].

2) Text Augmentation.: We utilize Homophone Substitution according to a dictionary comprising 8,000 Chinese characters grouped by their phonetic similarity (homophones). The Automatic Speech Recognition (ASR) output using iFlytek service [20] is traversed and each character will be replaced by one of its homophones with a small probability (e.g., 10%). This technique simulates the inaccuracies commonly encountered in ASR systems, thereby improving the robustness and generalization capabilities of the trained model.

C. Audio and Text Representations

1) Audio Feature Extraction.: The Wave2Vec feature extractor [21] is applied to extract acoustic features from the audio recordings by converting raw audio into spectral features. Audio recording is first converted into mono format, and normalized to improve model performance. We choose a maximum segment length of 40 seconds, which is based on observation of our audio segments for all phone recording samples. Segments longer than 40 seconds are truncated, while shorter ones are padded to maintain input consistency. The sampling rate is set to 8,000 Hz, and a sequence of 32,0000 values (an equivalence of 40 seconds) will be generated by Wave2Vec for each segment. This sequence is then passed to the HuBERT [22] encoder to generate an enhanced audio representation for the subsequent modeling steps. The above procedure generates a 999 sequence of 768 dimensional feature vectors.

Let $\mathbf{x}_{a_j} \in \mathbb{R}^{L_a \times d_a}$ denote the extracted feature vectors for segment a_j, where L_a is the sequence length (i.e., 999 in our current case), and $d_a = 768$ is the dimension of the acoustic feature vector. For each audio recording A_i, the extracted feature sequence can be represented as $\mathbf{X}_{A_i}$:

$$\mathbf{X}_{A_i} = [\mathbf{x}_{a_1}, \mathbf{x}_{a_2}, \ldots, \mathbf{x}_{a_{L_a}}].$$

2) Text Feature Extraction.: The transcriptions of the conversations in each audio recording segment is tokenized and encoded using a pre-trained RoBerta [23] encoder. We set the maximum token length L_t to 199 to ensure that the entire conversation or significant portions of it are captured within this limit. Text sequences longer than 199 are truncated, while shorter ones will be padded. Each token is embedded into a vector of dimension $d_t = 768$.

Similar to audio feature extraction, let $\mathbf{x}'_{a_j} \in \mathbb{R}^{L_t \times d_t}$ denote the extracted textual representation for the jth audio segment from A_i, we have:

$$\mathbf{X}'_{A_i} = [\mathbf{x}'_{a_1}, \mathbf{x}'_{a_2}, \ldots, \mathbf{x}'_{a_{L_t}}].$$

The feature extraction process for both audio and textual modalities is shown in Figure 1 Step 1.

D. Learning Complementary Signal from Audio

Although our original data modality is audio, it practically contains both textual and acoustic signals. Textual signal can be extracted through ASR, and usually its semantic meaning provides the main basis for audio classification. At the same time, acoustic signal can provide invaluable complementary information to the textual signal to enhance classification accuracy. In order to achieve this goal, we propose to first learn a complementary representation, and apply it to enhance the textual representation.

First, for each audio segment a_j within an audio recording A_i, we fuse $\mathbf{x}_{a_j}$ and $\mathbf{x}'_{a_j}$ using cross-attention encoder. Specifically, the cross-attention encoder first calculates the attention distribution between the query matrix $\mathbf{Q}_T$, the key matrix $\mathbf{K}_A$, and the value matrix $\mathbf{V}_A$ through the scaled dot-product:

$$\textbf{Attention}(\mathbf{Q}_T, \mathbf{K}_A, \mathbf{V}_A) = \text{softmax}\left(\frac{\mathbf{Q}_T \mathbf{K}_A^\top}{\sqrt{d_k}}\right)\mathbf{V}_A, \quad (1)$$

Where d_k is the dimension of the key matrix, and the softmax function is used to normalize the attention weight. This mechanism ensures that information from the audio modality A can be effectively extracted according to the contents of the text modality T. The output from the cross-attention module is input into a self-attention module, which further improves the expressiveness of the fused feature representation. In addition, a pointwise feed-forward Network is added to further enhance the feature representation after the self-attention mechanism.

At the same time, the textual sequence is independently input into the self-attention module to enhance its representation. As is shown in Figure1 step 2, we maintain a complementary fused channel and a separate textual channel as the input for bottleneck fusion.

Model	ACC_2	ACC_3	ACC_4	ACC_5
MMML	**78.09**	58.98	54.23	52.12
Ours	76.60	**63.83**	**61.70**	**60.28**

TABLE I
Experimental performances on the MarketCalls dataset. Acc is short for accuracy with the index number referring to the four different classification scenarios as described in Section IV-A.

E. Bottleneck Fusion

We apply a variant of the bottleneck fusion strategy [15] to fuse the audio-complementary signal with the textual signal. The core idea of bottleneck fusion is to significantly reduce the computational complexity while maintaining information exchange through the bottleneck layers (as shown in Figure 1). A set of bottleneck fusion vectors $\mathbf{T_{fsn}} = [\mathbf{T^1_{fsn}}, \mathbf{T^2_{fsn}}, \ldots, \mathbf{T^n_{fsn}}]$, where n denotes the number of bottleneck heads, is introduced to exchange information between the text modality features $\mathbf{T}$ and the text-audio fusion modality features $\mathbf{T_m}$. The input sequence can be expressed as:

$$\mathbf{T_z} = [\mathbf{T} \| \mathbf{T_{fsn}} \| \mathbf{T_m}]. \tag{2}$$

The information exchange and interaction between text features $\mathbf{T}$ and text-audio fusion features $\mathbf{T_m}$ is realized through the shared bottleneck layer $\mathbf{T_{fsn}}$. Specifically, the text features $\mathbf{T}$ and text-audio fusion features $\mathbf{T_m}$ are first concatenated with the current bottleneck representation $\mathbf{T_{fsn}}$ in the sequence dimension to form new inputs $[\mathbf{T}, \mathbf{T_{fsn}}]$ and $[\mathbf{T_m}, \mathbf{T_{fsn}}]$. Since the concatenation operation increases the sequence length, the system adjusts the attention mask accordingly to match the new input length to ensure that the self-attention mechanism can correctly handle the expanded sequence. Subsequently, the concatenated text channel $[\mathbf{T}, \mathbf{T_{fsn}}]$ and text-audio fusion channel $[\mathbf{T_m}, \mathbf{T_{fsn}}]$ are processed by the self-attention layer, as in Eq. 3 and 4 to capture the dependencies and importance between different positions, and their respective output features are generated through the attention output layer. Among the generated output features, the text feature $\mathbf{T}$ is extracted and updated from its output, and the representation of $\mathbf{T}$ is continuously optimized using the information fused by the bottleneck layer $\mathbf{T_{fsn}}$. At the same time, the bottleneck representation $\mathbf{T_{fsn}}$ is also updated through the output of the text-audio fusion channel to ensure that it can effectively integrate rich information from different modalities. Through this layer-by-layer dynamic update process, the bottleneck layer $\mathbf{T_{fsn}}$ not only promotes the continuous optimization of the text feature $\mathbf{T}$, but also enhances the collaborative processing capability of multimodal information, thereby significantly improving the overall expression ability and performance of the model.

We calculate the feature representation of each modality in a recursive manner as follows:

$$[\mathbf{T^{l+1}} \| \hat{\mathbf{T}}^{\mathbf{l+1}}_{\mathbf{fsn}}] = \text{Transformer}([\mathbf{T^l} \| \mathbf{T^l_{fsn}}]), \tag{3}$$

$$[\mathbf{T^{l+1}_m} \| \hat{\mathbf{T}}^{\mathbf{l+1}}_{\mathbf{fsn,m}}] = \text{Transformer}([\mathbf{T^l_m} \| \mathbf{T^l_{fsn}}]), \tag{4}$$

$$\mathbf{T^{l+1}_{fsn}} = \text{Avg}(\hat{\mathbf{T}}^{\mathbf{l+1}}_{\mathbf{fsn}}, \hat{\mathbf{T}}^{\mathbf{l+1}}_{\mathbf{fsn,m}}) \tag{5}$$

The benefits of our proposed bottleneck fusion mechanism include the following: 1) Shared Intermediate Representation. By introducing bottleneck representation, the bottleneck fusion mechanism establishes a shared information interaction bridge between different modalities, promoting information exchange and fusion between modalities. 2) Flexible Information Integration. The self-attention mechanism allows the model to dynamically adjust the importance of different modal information during the fusion process, capture complex dependencies, and thus achieve more flexible and effective information integration.

F. Contextual Modeling Using Bi-directional GRU

Each audio sample is divided into conversation segments, and these segments are serving as each other's contexts. Leveraging this contextual information for classification can improve the inference accuracy. We propose to use a bidirectional GRU (Bi-GRU) layer for context modeling.

The output representations from bottleneck fusion $\mathbf{T_j}$, $j = 1, 2, \ldots, l_i$ are provided as the input to the Bi-GRU layer. As is shown by Figure 1, each audio segment within an audio sample receives inputs from adjacent segments in both directions. The output of the GRU layer is denoted as $\mathbf{h}_n$, which encapsulates the final hidden states across all time steps in the input sequence.

G. Multi-task Learning

We employ a multi-task learning strategy to both accommodate business demands, as well as model generalizability and robustness. In multitask learning, the network backbone is shared among different related tasks as shown by Figure 1. All feature representations flowing through the network backbone are input into different task-specific output layers. These layers contain independent weight parameters, and will be updated independently. To optimize the performance of multi-task learning, each classification task computes its own cross-entropy loss, and the total loss fuction defined by Equation 6 is used to update the network parameters by backpropagation.

$$\begin{aligned}\mathcal{L}_{\text{total}} = &\mathcal{L}_{\text{five\_categories}} + \mathcal{L}_{\text{four\_categories}} + \\ &\mathcal{L}_{\text{three\_categories}} + \mathcal{L}_{\text{two\_categories}}\end{aligned} \tag{6}$$

IV. Experiments

A. Data and Baselines

The curated dataset MarketCalls consists of 1,000 phone call recordings made by sales representative to potential customers in various businesses using Mandarin. Only recordings that are longer than one minutes are selected and manually annotated by experienced salespersons from the associated marketing company. Five categories in purchase intent are created, including A - very positive,

c

Model	CMU-MOSI							CMU-MOSEI						
	ACC_{2Has0}	$F1_{Has0}$	ACC_{2Non0}	$F1_{Non0}$	ACC_7	MAE	Corr	ACC_{2Has0}	$F1_{Has0}$	ACC_{2Non0}	$F1_{Non0}$	ACC_7	MAE	Corr
LMF	-	-	82.5	82.4	33.90	0.917	0.695	-	-	82.0	82.1	48.00	0.623	0.700
TFN	-	-	80.8	80.7	34.90	0.901	0.698	-	-	82.5	82.1	50.20	0.593	0.677
MFM	-	-	81.7	81.6	34.50	0.877	0.706	-	-	84.4	84.3	51.30	0.568	0.703
MTAG	-	-	82.3	82.1	38.90	0.866	0.722	-	-	-	-	-	-	-
SPC	-	-	82.8	82.9	-	-	-	-	-	82.6	82.8	-	-	-
ICCN	-	-	83.0	83.0	39.00	0.862	0.714	-	-	84.2	84.2	51.60	0.565	0.704
MuLT	81.50	80.60	84.1	83.9	-	0.861	0.711	-	-	82.5	82.3	-	0.580	0.713
MISA	80.79	80.77	82.10	82.03	-	0.804	0.764	82.59	82.67	84.23	83.97	-	0.568	0.717
Self-MM	84.00	84.42	85.98	85.95	-	0.713	0.798	85.17	82.81	82.53	85.17	85.30	0.530	0.765
MAGBERT	84.20	84.10	86.10	86.00	-	0.712	0.796	84.70	84.50	-	-	-	-	-
MIMIM	84.14	84.00	86.06	85.98	46.65	0.700	0.800	82.24	82.66	85.97	85.94	54.24	0.526	0.772
TEASEL	84.79	84.72	87.5	85	47.52	0.644	0.836	-	-	-	-	-	-	-
UniMSE	85.85	85.83	86.9	86.42	48.68	0.691	0.809	85.86	85.79	87.5	87.46	54.39	0.523	0.773
MMML	85.91	85.85	88.16	88.15	48.25	0.6429	0.838	86.32	86.23	86.73	86.49	54.95	0.5174	0.7908
MMML(context)	**87.51**	**87.45**	89.69	89.67	50.34	**0.5831**	**0.8693**	**87.24**	**87.18**	88.02	88.15	**55.74**	**0.4922**	**0.8137**
OURS	86.15	86.14	**91.21**	**91.38**	**51.02**	0.6458	0.8243	87.16	87.02	**88.62**	**88.85**	55.58	0.5061	0.7972

TABLE II
Experimental Performances on the CMU-MOSI and CMU-MOSEI datasets.

Model	ACC7	F17	surprise	anger	sadness	neutal	joy
CTNET	-	60.5	52.7	44.6	32.5	77.4	56.0
DF-ERC	**68.28**	**67.03**	60.27	55.50	**43.89**	80.17	**65.93**
OURS	64.82	63.19	**72.24**	**55.96**	35.78	**85.12**	56.98

TABLE III
Experimental performances on the MELD dataset.

B - neutral and receptive, C - a little impatient and negative, D - explicit refusal, and E - no intent or not relevant. After data augmentation in audio and textual signals separately, we obtained a total number of 7042 and 4586 audio recording samples respectively for model training and evaluation.

In addition to the five-categories task, we create three other classification tasks by combining the above five categories into the following scenarios: four-categories (A, B, C, D & E); three-categories (A, B & C, D & E); and two-categories (A & B & C, D & E).

To fully evaluate our proposed method, we also experiment with several popular open benchmark datasets, including the CMU-MOSI [1] and MOSEI datasets [2], and the MELD dataset [3]. We choose recent state-of-the-art baselines in multimodal emotion recognition and sentiment analysis, including DF-ERC [24] and MMML [13]. Other baselines are shown in various tables when appropriate.

B. Hyper-parameter Tuning and Implementation

We adopt a grid-search strategy in hyper-parameter tuning and obtain the following parameters for all experiments: learning rate is 1e-5, number of hidden layers is 4, number of accumulation steps in gradient is 4, number of bottleneck layers is 2 with 4 bottleneck nodes in each layer, number of GRU layers is 2 with 128 nodes in each layer, L2 regularization is selected and the dropout rate is 0.3.

V. Results

Type	specific type	ACC_2	ACC_3	ACC_4	ACC_5
AUG	audio-augment	**76.60**	**63.83**	**61.70**	60.28
	text-augment	72.34	58.16	60.99	**60.99**
	no-augment	75.89	60.02	58.16	57.46
PREPR	Rm-silence	64.61	46.88	46.10	45.39
	Kp-silence	**76.60**	**63.83**	**61.70**	**60.28**
MODEL	no-BF+MTL	73.05	55.32	54.89	56.74
	no-MTL	71.23	53.90	52.48	51.77
	full	**76.60**	**63.83**	**61.70**	**60.28**

TABLE IV
Ablation Analyses on MarketCalls dataset. 1) AUG: augmentation strategies applied on audio (audio-augment) and text (text-augment) modalities. 2) PREPR: removing (Rm-silence) or keeping (Kp-silence) silence in original audio samples. 3) MODEL: evaluation on the impacts of bottleneck fusion and multi-task learning. Bottleneck fusion and multi-task learning are removed in no-BF+MTL, while only multi-task learning is removed in no-MTL.

1) Performance Comparisons: Table I, II and III show the experimental results on the MarketCalls, CMU-MOSI and CMU-MOSEI, and MELD datasets, respectively. We can see that in the MarketCalls dataset, MSMT-FNet obtains better performance in three out of four classification tasks when compared to MMML; in the CMU-MOSI and CMU-MOSEI datasets, MSMT-FNet obtains better or comparable results in most evaluation metrics when compared to MMML; and in the MELD dataset, MSMT-FNet obtains comparable performance when compared to DF-ERC.

We acknowledge that no single method including the existing state-of-the-art baselines has attained consistently best performance across all evaluation metrics in all four datasets. However, our proposed network achieves comparable performances on the three open benchmark datasets in sentiment inferences, which suggests that even though our proposed method is designed mainly for customer

intent classification, it can still be generalized toward a different domain.

2) Ablation Analyses: Table IV shows the ablation analyses on the MarketCalls dataset with MSMT-FNet. Augmenting data using the audio modality leads to best performance in all tasks. Keeping the silence portions within the audio samples results in significantly better performance than removing them. This could be due to the potential signal encoded by silence itself with regard to customers' purchasing intents. Some interaction effects are observed between bottleneck fusion and multi-task learning, given that the full model obtains the best performance, while leaving out only the multi-task learning component results in worst performance.

VI. Conclusion

In this work, we propose a multi-segment multitask fusion network that has been shown to be effective in both the MarketCalls dataset, and three open benchmark datasets. Few existing research has been conducted in this increasingly important application in the field of telemarketing, as more and more commercial activities are moving from offline to online spaces. Our contributions include both proposing a novel solution as well as contributing a new benchmark dataset to the field.

References

[1] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency, "Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016.

[2] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Iryna Gurevych and Yusuke Miyao, Eds., Melbourne, Australia, July 2018, pp. 2236–2246, Association for Computational Linguistics.

[3] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," 2019.

[4] Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Vikram Singh, and Ashutosh Modi, "Cogmen: Contextualized gnn based multimodal emotion recognition," 2022.

[5] Shiqing Zhang, Yijiao Yang, Chen Chen, Xingnan Zhang, Qingming Leng, and Xiaoming Zhao, "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," Expert Systems with Applications, vol. 237, pp. 121692, 2024.

[6] Huiting Fan, Xingnan Zhang, Yingying Xu, Jiangxiong Fang, Shiqing Zhang, Xiaoming Zhao, and Jun Yu, "Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals," Information Fusion, vol. 104, pp. 102161, 2024.

[7] Garima Sharma, Kartikeyan Umapathy, and Sridhar Krishnan, "Trends in audio signal feature extraction methods," Applied Acoustics, vol. 158, pp. 107020, 2020.

[8] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li, "Learning alignment for multimodal emotion recognition from speech," 2020.

[9] Lucas Goncalves and Carlos Busso, "Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features," IEEE Transactions on Affective Computing, vol. 13, no. 4, pp. 2156–2170, 2022.

[10] Zheng Lian, Bin Liu, and Jianhua Tao, "Ctnet: Conversational transformer network for emotion recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 985–1000, 2021.

[11] Bo Yang, Bo Shao, Lijun Wu, and Xiaola Lin, "Multimodal sentiment analysis with unidirectional modality translation," Neurocomputing, vol. 467, pp. 130–137, 2022.

[12] Kyeonghun Kim and Sanghyun Park, "Aobert: All-modalities-in-one bert for multimodal sentiment analysis," Information Fusion, vol. 92, pp. 37–45, 2023.

[13] Zehui Wu, Ziwei Gong, Jaywon Koo, and Julia Hirschberg, "Multimodal multi-loss fusion network for sentiment analysis," 2024.

[14] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Tensor fusion network for multimodal sentiment analysis," 2017.

[15] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun, "Attention bottlenecks for multimodal fusion," in Advances in Neural Information Processing Systems, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, Eds. 2021, vol. 34, pp. 14200–14213, Curran Associates, Inc.

[16] Michael Crawshaw, "Multi-task learning with deep neural networks: A survey," 2020.

[17] Bagus Tris Atmaja, Mifta Nur Farid, and Dhany Arifianto, "Speech enhancement on smartphone voice recording," in Journal of Physics: Conference Series. IOP Publishing, 2016, vol. 776, p. 012072.

[18] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition.," in Interspeech, 2015, vol. 2015, p. 3586.

[19] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," arXiv preprint arXiv:1904.08779, 2019.

[20] "iflytek asr services," 2024, Accessed: September 12, 2024.

[21] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[22] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021.

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[24] Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li, "Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition," in Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 5923–5934.