Cross-transfer Knowledge between Speech and Text Encoders to Evaluate Customer Satisfaction

Luis Felipe Parra-Gallego[1,2], *Tilak Purohit*[3,4], *Bogdan Vlasenko*[3], *Juan Rafael Orozco-Arroyave*[1,5], *Mathew Magimai.-Doss*[3]

[1]GITA Lab. University of Antioquia, Medellín, Colombia. [2]Konecta Group S.A.S. Medellín, Colombia. [3]Idiap Research Institute, Martigny, Switzerland.
[4]EPFL, École polytechnique fédérale de Lausanne, Switzerland.
[5]Pattern Recognition Lab. Friedrich Alexander University, Erlangen-Nuremberg

lfelipe.parra@udea.edu.co, {tilak.purohit, bogdan.vlasenko}@idiap.ch

Abstract

Customer Satisfaction (CS) in call centers influences customer loyalty and the company's reputation. Traditionally, CS evaluations were conducted manually or with classical machine learning algorithms; however, advancements in deep learning have led to automated systems that evaluate CS using speech and text analyses. Previous studies have shown the text approach to be more accurate but relies on an external ASR for transcription. This study introduces a cross-transfer knowledge technique, distilling knowledge from the BERT model into speech encoders like Wav2Vec2, WavLM, and Whisper. By enriching these encoders with BERT's linguistic information, we improve speech analysis performance and eliminate the need for an ASR. In evaluations on a dataset of customer opinions, our methods achieve over 92% accuracy in identifying CS categories, providing a faster and cost-effective solution compared to traditional text approaches.

Index Terms: Customer satisfaction, cross-transfer knowledge, Spoken Language Understanding

1. Introduction

Customer satisfaction (CS) is a relevant metric in call centers, serving as an indicator of the matching between customer expectations and products, services, and customer experience provided by the company [1]. CS directly influences customer loyalty, retention, and the overall reputation of the organization. This implies that a company grows as satisfied customers tend to make more purchases and recommend products and services to other people, thereby attracting more customers [2]. Therefore, assessing CS is essential for companies aiming to enhance their quality of service (QoS). Typically, CS evaluations were conducted manually by QoS experts who would select and analyze a small sample of call or voicemail recordings. However, the recent advancements in deep learning (DL) have led to the development of automated methods for CS assessment, focusing on two primary approaches: speech and text analysis.

Some studies have explored the use of speech representations for CS analysis, dividing these representations into two main categories: emotion-oriented/knowledge-based and data-driven approaches. Emotion-oriented features aim to model emotions that influence CS, encompassing different speech dimensions such as prosody, articulation, voice-quality, spectral characteristics, and patterns of silence and pause [3, 4, 5]. Conversely, data-driven approaches rely on algorithms to learn representations directly from the data, producing general yet complex speech features suitable for various tasks. Wav2Vec [6, 7] stands out as a notable data-driven technique applied in CS modeling [8, 9, 10]. Other popular approaches for speech processing tasks include WavLM [11] and HuBERT [12]. Despite the advancements in speech-based methods, evidence suggests that text-based analysis often yields higher accuracy in CS evaluation [13, 14].

Recent improvements in automatic speech recognition (ASR) and natural language understanding (NLU) systems have enhanced the reliability of text analysis. Also known as Spoken Language Understanding (SLU), this method aims to directly extract meaning or intent from spoken utterances [15]. Conventionally, its pipeline comprises two main steps. First, an ASR system is responsible for converting a spoken utterance into a text transcript. Then, the transcripts are processed by an NLU system intended to model CS concepts. Various studies have demonstrated the effectiveness of NLU systems based on text features like such as TF-IDF [16], Word2Vec [17], and BERT [18] for modeling CS concepts [9, 19, 20]. However, using these traditional SLU approaches in call center applications presents three main challenges. Firstly, poor-quality speech transcripts can adversely affect performance in downstream tasks, as demonstrated empirically in [21, 22]. Secondly, relying on an external ASR for text transcription introduces additional complexity for modeling, thereby increasing the required inference processing time. Finally, call recordings often contain sensitive information about the customers, posing a risk of exposure during the transcription process [23]. Hence, implementing new mechanisms to protect the privacy of the speaker is essential.

In response to these challenges, end-to-end (E2E) SLU systems have emerged as a pivotal technology. In this new scenario, a single trainable model can directly model semantic patterns from a spoken utterance, eliminating the need for producing a text transcription [24, 25]. Consequently, the model becomes more compact and can be fully optimized directly on the targeted metric for the downstream task. This makes the E2E approach practical for industrial applications where optimal use of computational resources is crucial. Despite widespread adoption in intent classification, E2E SLU systems remain under-investigated in CS evaluation.

This paper compares traditional methods based on unimodal and multimodal approaches to classify satisfied vs. unsatisfied customers using voicemails from a call center dataset. For the speech modality, we explore three data-driven feature encoders: Wav2Vec2, WavLM, and the Whisper encoder [26]. For text analysis, we employ BERT representations to capture linguistic cues. In the multimodal approach, we combine the optimal features from each modality using late and early fusion, as well as Gated Multimodal Units (GMU), which are based on DL and perform information fusion at an intermediate

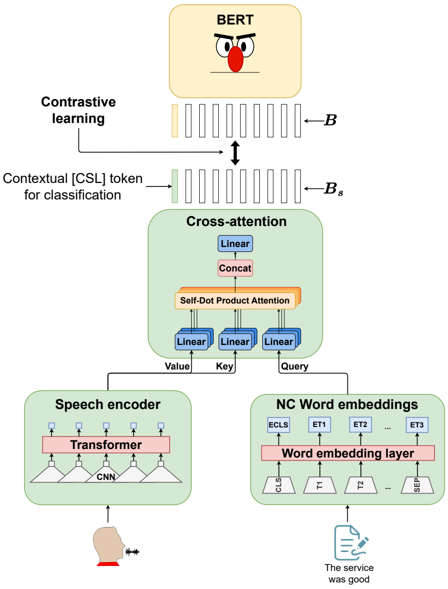

Figure 1: *General methodology followed in this study. Figure adapted from [24].*

level [27]. We then introduce an E2E SLU system to directly model CS patterns from speech. This involves developing a teacher-student architecture to transfer knowledge from BERT to the three speech encoders. This process generates linguistically enriched speech representations, which are subsequently utilized for classification. The key contributions of our study are summarized as follows:

1. The E2E SLU framework, which is underexplored in CS tasks, is evaluated in spoken opinions. This approach highlights the benefits of eliminating ASR systems, simplifying the process, and protecting the privacy of customer voicemails.
2. Typical works in English use linguistically rich datasets for pre-training [24]; however, this is not the case for Spanish. Therefore, we evaluate three state-of-the-art (SOTA) pre-trained speech encoders. Using these encoders allows us to focus pre-training on specific layers, such as the non-contextual (NC) embedding layer and the cross-attention layer, using low-resource Spanish data.
3. By leveraging cross-transfer knowledge, our approach achieves accuracies of up to 92%, and it outperforms traditional text-based methods by being five times faster in terms of inference processing.

2. Methodology

The methodology is divided into two main steps: cross-transfer knowledge[1] and classification. Figure 1 illustrates the methodology followed in this study [2].

2.1. Cross-transfer knowledge

This study employs a teacher-student architecture to distill knowledge from BERT into the three different data-driven speech encoders, as described below.

Speech encoder: To extract speech embeddings, we employ three SOTA systems in speech processing tasks: Wav2Vec2, WavLM, and the Whisper encoder. The first two encoders are pre-trained using self-supervised learning (SSL) techniques. These models are designed to capture general representations from speech. Whisper is a task-specific model trained for ASR tasks, making it valuable for extracting speech representations with linguistic content.

NC word embeddings: It is essentially a lookup table that maps each word to a vector representation. In this study, this layer is initialized by duplicating the word embedding layer pre-trained on BERT. It mimics the structure found in BERT preceding the stack of transformers. Replicating the NC layer ensures consistency between the teacher and student models by maintaining identical sequence dimensions. Furthermore, this replication simplifies the alignment process, as the student is already familiar with aspects of the teacher, allowing it to focus mainly on the transformer stack, which models contextual information.

BERT: This model is a SSL language model based on transformer layers, designed to extract meaningful linguistic representations. For this study, we utilizes BETO [28] as the teacher model, which follows BERT's model architecture but was specifically trained on a Spanish dataset.

Cross-attention mechanism : In our study, we implement a standard attention mechanism [29, 30], assigning speech encoder outputs as both key and value, while text outputs as the query. This allows text to obtain contextual information from speech modality. The formula used is:

$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{d_k}}\right)\boldsymbol{V},$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are the matrices for query, key, and value, with d_k as the feature dimension.

Contrastive loss: Once the contextual representation $\boldsymbol{B_s}$ is generated, it is then aligned with the semantically rich BERT contextual representation $\boldsymbol{B}$ on a token-by-token basis, given that both shared the same sequence length n. To accomplish this, we employ the tokenwise contrastive loss, as proposed in [24], which ensures alignment of the representations at the token level. The contrastive loss is defined as:

$$\mathcal{L}_{con} = -\frac{\tau}{2b}\sum_{i=1}^{b}\log\frac{\exp(s_{ii})}{\sum_{j=1}^{b}\exp(s_{ij})} + \log\frac{\exp(s_{ii})}{\sum_{j=1}^{b}\exp(s_{ji})}$$

where s_{ij} represents the cosine similarity between rows i and j in $\boldsymbol{B}$ and $\boldsymbol{B_s}$, and τ is a temperature hyperparameter of the cosine similarity function.

2.2. Classification stage

After performing cross-transfer knowledge, we can train a classifier on the downstream task using BERT-like speech features. At this stage, transcripts of the recordings are unavailable. Therefore, the NC [CLS] learned during pre-training is processed through the cross-attention layer to attend over the speech sequence. This procedure allows capturing context directly from speech, generating a contextual BERT-like token [CLS] as a sentence-level representation for classification.

3. Experiments

This section presents information about the KONECTADB dataset, provides details of the implementation, and introduces

[1] Here, the terms 'cross-transfer knowledge', 'knowledge distillation', and 'pre-training' are used interchangeably.

[2] Code available at: `https://github.com/lfelipeparra/cross-transfer-knowledge`

the baseline methods. Subsequently, the results obtained are discussed.

3.1. Dataset

KONECTADB [5] is employed to evaluate CS in a real-world scenario. The database comprises spoken customer opinions (voicemails) recorded at the end of conversations with agents at the Konecta call center. In these voicemails, customers gave spontaneous evaluations of the quality of the service provided by the agent. Before recording the voicemails, customers were informed that their speech was going to be recorded. All participants were adults who were native speakers of Colombian Spanish. The recordings were captured at 8kHz with 16-bit resolution. A total of 2364 recordings were collected and annotated by QoS experts from Konecta. They listened to and evaluated the voicemails, labeling whether the customers were satisfied or not. The recordings were automatically transcribed by Whisper `large-v3`. Table 1 describes the data distribution for KONECTADB. Gender balance was assured by a chi-square test with $p \approx 1$.

Table 1: *Data distribution and general information for the KONECTADB*

	Dissatisfied	**Satisfied**
Number of samples	1259	1105
Duration ($\mu \pm \sigma$)	34±23 s	16 ±11 s
Number of male	711	532
Number of female	548	573

3.2. Implementation details and metrics

The proposed method, described in Section 2, is evaluated on KONECTADB. The dataset follows a bootstrapping strategy of 80% for training, 10% for validation, and 10% for testing. Due to time constraints, we do not employ nested cross-validation. This is because pre-training the speech encoders and training the classifier for each fold are time-consuming processes. Speech features are extracted from the voicemails recordings while both NC word embeddings and BERT representations are computed from text transcripts. Due to the difference in feature dimensions between both representations, we process the speech features through a fully-connected layer to match the dimension d_k outlined in Subsection 2.1, Cross-transfer knowledge.

In the cross-transfer knowledge stage, we use the speech-text pairs from the training split on KONECTADB. Here, the validation split is utilized for model selection. The experiment is conducted on an NVIDIA RTX 3090 GPU for 200 epochs. We employ a batch size of 64 voicemails and the AdamW optimizer with a learning rate set at 1e-4. The temperature hyperparameter τ is adjusted to 0.07.

In the classification stage, an SVM classifier with a radial basis function (RBF) kernel is used to classify contextual BERT-like representations. The hyperparameters were optimized through grid search for $C \in \{10^0, 10^1, \ldots, 10^3\}$ and $\gamma \in \{10^{-6}, 10^{-5}, \ldots, 10^2\}$, using logarithmic steps. The model optimization is conducted on an Intel(R) Core(TM) i7-7700K CPU @ 4.20GHz processor. Similar to the cross-transfer knowledge stage, the model is trained on the training split, with the validation split used for hyperparameter optimization and model selection. Area Under ROC curve (AUC), Accuracy (ACC), Sensitivity (SEN) and Specificity (SPE) are computed over the test split.

3.3. Baseline methods

We conduct speech and text analyses using Wav2Vec2, WavLM, Whisper, and BERT representations for CS classification in unimodal settings. For the multimodal approach, we select the best representations from each modality and apply early and late fusion strategies, as well as the one based on GMU [27] model, to merge speech and text. Late fusion employs a weighted averaging approach, with weights being optimized through grid search. The search space for these weights ranges from 0.1 to 0.9, incrementing in steps of 0.1. The GMU is jointly optimized with DL classifiers by minimizing the objective function. Two classification methods are evaluated: SVM and DL. For the SVM classifier, we use the settings defined in Subsection 3.2, except that the input consists of a static representation obtained by averaging the embeddings. For the DL classifier, we explore two contextual layers: a bidirectional LSTM (128 units per direction) and a self-attention layer (scaled-dot product method). Their contextual outputs are globally averaged and fed into a classifier comprising a 128-unit ReLU-activated hidden layer, followed by a Softmax classification layer. To ensure consistency with other studies utilizing the same database, both classifiers are evaluated following a 5-fold cross-validation approach. In each fold, 10% of the training split is randomly selected for hyperparameter optimization and model selection. To measure performance, the predictions of the test split from each fold (as well as the ground truth labels) are concatenated. Subsequently, metrics are calculated across the entire dataset.

3.4. Results and discussion

Baseline: The results of baseline methods for both unimodal and multimodal analyses are presented in Table 2. Results from methods detailed in [5], denoted with an asterisk (*), are also included for comparative analysis. In the speech scenario, all considered representations outperform previous studies, improving by about 10% in absolute accuracy when comparing Whisper to Wav2Vec*. We believe the improvement can be attributed to three factors: (1) Wav2Vec2, fine-tuned on a Spanish corpus, is better at modeling speech patterns in Spanish and its output includes linguistic information due to training on ASR tasks. (2) WavLM, trained under challenging acoustic conditions such as noisy/overlapping recordings, is well-suited for real-world data like KONECTADB. (3) The Whisper encoder is effective in representing speech features with linguistic content, as it is a SOTA model trained on multilingual ASR tasks. In the text scenario, BERT emerges as the best-performing method after fine-tuning, achieving an accuracy of 94.25%. The nature of the task, in which customers might linguistically express positive/negative opinions without exhibiting any specific emotional traits while speaking, makes text analysis is more reliable than speech analysis. That is why the application of text models proves to be more accurate than acoustic analysis. In multimodal analysis, fusion techniques do not significantly enhance classification performance, with an absolute improvement of 0.16% when comparing *late(*Whisper, BERT$_{ft}$*)* to BERT$_{ft}$. This suggests that acoustic representations may not offer complimentary information to text, which already performs very well. Moreover, the increased model complexity makes the optimization process harder.

Cross-transfer knowledge: The results are presented in Table 3. Note that we also measure the inference time required to process a sample of 30 seconds long for each model. This means that the time required by Whisper `large-v3` is also

Table 2: *Comparison of different unimodal and multimodal approaches on KONECTADB. $*$ indicates the methods emerged from [5]. The results are given in [%]. ft suffix means fine-tuning.*

Feature representation	**Model**	**AUC**	**ACC**	**SEN**	**SPE**
Speech features					
Articulation *	SVM	-	73,50	75,90	70,80
xvector *		-	66,30	70,60	61,30
I2012PC *		-	74,20	79,50	68,10
WavLM		90,89	82,90	84,42	81,18
Wav2Vec *	BiLSTM	-	77,40	80,10	74,60
Wav2Vec2		88,14	80,23	83,71	76,28
WavLM		90,48	82,61	82,76	82,44
Whisper		93,35	86,75	88,32	84,98
Wav2Vec2	Attention	89,39	82,22	87,67	76,01
WavLM		91,98	84,00	87,36	80,18
Whisper		**96,90**	**88,56**	**89,65**	**87,32**
Text features					
Word2Vec *	SVM	-	87,90	89,25	86,90
BERT		96,06	89,25	89,59	88,86
Word2Vec *	BiLSTM	-	90,90	92,60	89,20
BERT		95,67	89,67	89,11	90,22
BERT	Attention	97,89	92,76	94,43	90,86
BERT$_{ft}$		**98,72**	**94,25**	**95,47**	**92,85**
Multimodal					
*early(*Artic, Word2Vec*)* *	SVM	-	87,80	88,90	86,60
*early(*Whisper, BERT*)*		97,38	92,29	93,32	91,13
*early(*Whisper, BERT$_{ft}$*)*		98,30	94,28	94,75	93,75
*late(*Whisper, BERT*)*		97,62	92,17	93,08	91,13
*late(*Whisper, BERT$_{ft}$*)*		**98,49**	**94,41**	**94,91**	**93,84**
*intern(*Wav2Vec, Word2Vec*)* *	GMU	-	90,80	92,40	89,30
*intern(*Whisper, BERT*)*		97,23	91,96	93,16	90,49
*intern(*Whisper, BERT$_{ft}$*)*		98,37	94,07	94,83	93,21

considered for the traditional text analysis. Overall, speech representations show improvement after applying knowledge distillation from the BERT model when compared to non-distilled baseline counterparts. Furthermore, all distilled models obtain an additional gain of about 2 percentage points in ACC when a specialized teacher (BERT$_{ft}$) is used. This indicates that the robustness of BERT models is successfully transferred to speech encoders, making them more efficient in modeling spoken opinions.

The BERT$_{ft}$-like Whisper model outperforms other distilled representations across all performance metrics. However, this model is more complex than its counterparts, requiring roughly twice the time to process a 30s sample. Both distilled Wav2Vec2 and WavLM models show similar performance in terms of ACC, being Wav2Vec2 faster by 3 seconds for the same sample.

One of the most significant advantages of using distilled models is their efficiency in terms of inference processing time. Both BERT and BERT$_{ft}$ representations need 212.33 seconds to process a 30s spoken opinion, which is considerably higher than that of any of the listed speech encoders. For instance, the BERT$_{ft}$-like Whisper can process the same sample about five times faster (in just 42.96 seconds), offering a much more efficient solution without substantially compromising performance. This type of system also eliminates the need for transcribing spoken utterances, a crucial advantage in safeguarding private and sensitive customer information, which could be compromised by exposure to transcription data. These results suggest that the introduced systems are ideal for call center applications where inference processing time is critical to timely evaluate and improve the quality of service and reduce usage costs while keeping sensitive information secure.

Table 3: *Speech encoder performance before and after applying cross-transfer knowledge. ft: fine-tuned model.* ***Time:*** *inference processing time in sec. for a sample of 30s.*

Feature	**AUC**	**ACC**	**SEN**	**SPE**	**Time**
Wav2Vec2	89,14	82,45	86,90	77,37	23,06
WavLM	89,89	81,81	84,12	79,18	26,41
Whisper encoder	96,96	88,58	93,25	83,26	42,01
BERT	96,24	90,06	92,46	87,33	212,33
BERT$_{ft}$	97,20	94,29	95,63	92,76	212,33
BERT-like Wav2Vec2	92,66	84,35	88,88	79,18	24,01
BERT$_{ft}$-like Wav2Vec2	91,98	86,05	92,06	79,19	**24,01**
BERT-like WavLM	91,08	85,20	86,90	83,26	27,36
BERT$_{ft}$-like WavLM	92,48	85,84	88,89	82,35	27,36
BERT-like Whisper	97,03	90,70	94,05	86,88	42,96
BERT$_{ft}$-like Whisper	**98,06**	**92,60**	**94,44**	**90,49**	42,96

Table 4: *Feature importance scores for late fusion and GMU strategies. s_{W} - the importance score for Whisper. s_{B} - the importance score for BERT or BERT-like, when applicable.*

Combinations	**Late Fusion**		**GMU**	
	s_{W}	s_{B}	s_{W}	s_{B}
Whisper + BERT	0.50	0.50	0.02	0.98
Whisper + BERT$_{ft}$	0.34	0.65	0.00	1.00
Whisper + BERT-like	0.43	0.57	0.00	1.00
Whisper + BERT$_{ft}$-like	0.22	0.78	0.33	0.67

Feature importance analysis: To confirm the potential of the new linguistically enriched features for modeling CS, we perform a feature importance analysis by merging the speech representations with the generated BERT-like features. We hypothesize that late fusion and GMU strategies would favor BERT-like features. In late fusion, importance scores are obtained from weights optimized in the validation set. For the GMU approach, as described in [27], we average the gate vector z generated in each test sample to determine the significance of each modality. Here, important scores correspond to the portion of samples in the test set that lean towards the language modality ($z <= 0.5$) and how many towards the speech ($z > 0.5$). Table 4 shows the importance scores for Whisper, BERT, and BERT-like in each method. Generally, BERT-like features carry more weight than original speech features. In the late fusion context, the incorporated features exhibit a trend consistent with the findings related to BERT. A similar pattern is observed in the GMU configuration, where the model is almost entirely focused on BERT-like features, except for BERT$_{ft}$-like, which demonstrates an importance value of 0.67.

4. Conclusion

This research explored the efficacy of linguistically enriched speech features for CS evaluation in call center environments, introducing a novel approach that leverages cross-transfer knowledge and classification techniques. Our methodology involved a teacher-student architecture, where BERT serves as the teacher to enrich speech encoder representations. These enriched representations were then evaluated for their classification performance. Our experiments demonstrated that SOTA speech encoders, such as Wav2Vec2, WavLM, and Whisper, when enhanced with BERT's linguistic capabilities, outperformed traditional speech-only models in CS evaluation tasks, with the Whisper encoder yielding better performance. BERT-like Whisper representations achieved approximately 92% accuracy in distinguishing between satisfied and dissatisfied customers, providing shorter inference processing times compared to conventional text analysis techniques. The feature importance analysis further confirmed our hypothesis that BERT-like features are more critical for accurate CS evaluation than traditional speech features. This was confirmed in both, late fusion and GMU strategies, where BERT-like features consistently carried more importance.

5. Acknowledgments

This work was partially funded by (a) CODI at UdeA grant # IAPFI23-1-01, (b) the Swiss National Science Foundation through the Bridge Discovery project EMIL: Emotion in the loop - a step towards a comprehensive closed-loop deep brain stimulation in Parkinson's disease (grant no. 40B2–0_194794 EMIL), and (c) the Swiss Innovation Agency Innosuisse through the flagship project IICT grant No PFFS-21-47.

6. References

[1] J. Dietz, "Satisfaction: A behavioral perspective on the consumer," *Journal of Consumer Marketing*, vol. 14, no. 4-5, pp. 401–404, 1997.

[2] J. McColl-Kennedy and U. Schneider, "Measuring customer satisfaction: why, what and how," *Total quality management*, vol. 11, no. 7, pp. 883–896, 2000.

[3] A. Ando *et al.*, "Customer satisfaction estimation in contact center calls based on a hierarchical multi-task model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 715–728, 2020.

[4] W. Han *et al.*, "Ordinal learning for emotion recognition in customer service calls," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2020, pp. 6494–6498.

[5] L. F. Parra-Gallego and J. R. Orozco-Arroyave, "Classification of emotions and evaluation of customer satisfaction from speech in real world acoustic environments," *Digital Signal Processing*, vol. 120, p. 103286, 2022.

[6] S. Schneider *et al.*, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. INTERSPEECH*, pp. 3465–3469.

[7] A. Baevski *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[8] N. Lackovic *et al.*, "Healthcall corpus and transformer embeddings from healthcare customer-agent conversations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[9] M. Macary *et al.*, "On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 373–380.

[10] T. Deschamps-Berger, L. Lamel, and L. Devillers, "Investigating transformer encoders and fusion strategies for speech emotion recognition in emergency call center conversations." in *International Conference on Multimodal Interaction*, 2022, pp. 144–153.

[11] S. Chen *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[12] W.-N. Hsu *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[13] J. Luque *et al.*, "The role of linguistic and prosodic cues on the prediction of self-reported satisfaction in contact centre phone calls." in *Proc. INTERSPEECH*, 2017, pp. 2346–2350.

[14] M. Macary *et al.*, "Acoustic and linguistic representations for speech continuous emotion recognition in call center conversations," *arXiv preprint arXiv:2310.04481*, 2023.

[15] G. Tur and R. De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons, 2011.

[16] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[17] T. Mikolov *et al.*, "Efficient estimation of word representations in vector space," in *International Conference on Learning Representations (ICLR)*, 2013.

[18] J. Devlin *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019.

[19] Á. Aldunate *et al.*, "Understanding customer satisfaction via deep learning and natural language processing," *Expert Systems with Applications*, vol. 209, p. 118309, 2022.

[20] Y. Kim, J. Levy, and Y. Liu, "Speech sentiment and customer satisfaction estimation in socialbot conversations," in *Proc. INTERSPEECH*, 10 2020, pp. 1833–1837.

[21] G. Saon, B. Ramabhadran, and G. Zweig, "On the effect of word error rate on automated quality monitoring," in *IEEE Spoken Language Technology Workshop*, 2006, pp. 106–109.

[22] Y. Park *et al.*, "An empirical analysis of word error rate and keyword error rate." in *Proc. INTERSPEECH*, vol. 2008, 2008, pp. 2070–2073.

[23] M. Tang, D. Z. Hakkani-Tür, and A. GokhanTur, "Preserving privacy in spoken language databases," in *Proc. International Workshop on Privacy and Security Issues in Data Mining*, 2004.

[24] V. Sunder *et al.*, "Tokenwise contrastive pretraining for finer speech-to-bert alignment in end-to-end speech-to-intent systems," in *Annual Conference of the International Speech Communication Association*, 2022.

[25] L. Lugosch *et al.*, "Speech Model Pre-Training for End-to-End Spoken Language Understanding," in *Proc. INTERSPEECH*, 2019, pp. 814–818.

[26] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning (ICML)*, 2023, pp. 28 492–28 518.

[27] J. Arevalo *et al.*, "Gated multimodal units for information fusion," in *International Conference on Learning Representations (ICLR)*, 2017.

[28] J. Cañete *et al.*, "Spanish pre-trained bert model and evaluation data," in *International Conference on Learning Representations (ICLR)*, 2020.

[29] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations (ICLR)*, 2015.

[30] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.