# A DIALOGICAL EMOTION DECODER FOR SPEECH EMOTION RECOGNITION IN SPOKEN DIALOG

*Sung-Lin Yeh, Yun-Shao Lin, Chi-Chun Lee*

Department of Electrical Engineering, National Tsing Hua University, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

## ABSTRACT

Developing a robust emotion speech recognition (SER) system for human dialog is important in advancing conversational agent design. In this paper, we proposed a novel inference algorithm, a dialogical emotion decoding (DED) algorithm, that treats a dialog as a sequence and consecutively decode the emotion states of each utterance over time with a given recognition engine. This decoder is trained by incorporating intra- and inter-speakers emotion influences within a conversation. Our approach achieves a $70.1\%$ in four class emotion on the IEMOCAP database, which is $3\%$ over the state-of-art model. The evaluation is further conducted on a multi-party interaction database, the MELD, which shows a similar effect. Our proposed DED is in essence a conversational emotion rescoring decoder that can also be flexibly combined with different SER engines.

***Index Terms***— speech emotion recognition, conversation, dialogical emotion decoder

## 1. INTRODUCTION

Emotion is a fundamental internal state that affects the way humans behave and interact with one other [1]. Thanks to the advancement in deep learning techniques, many researches have made use of neural nets in achieving promising performances in speech emotion recognition (SER) and helping the design of emotion-aware solutions. Examples of deep learning algorithms for SER include the use of convolutional neural networks (CNN), recurrent neural networks (RNN) with attention [2, 3, 4] and generative adversarial networks (GAN) [5]. Recently, as the development in realizing conversational agent has become more prevalent, the ability to computationally model emotion during human conversation is becoming the next critical step. It leads not only to a better understanding of human's conversational flow but further enables a better design of emotion-aware spoken dialog system. This direction of research has, in fact, led to an increasing interest in computationally modeling emotion recognition in conversations (ERC) [6]. Unlike straightforward isolated utterance-based emotion recognition, ERC requires proper handling of the dialog's contextual history, e.g., interlocutors mutual influence and self transition through time.

Several research works have proposed approaches in modeling such a conversational context for SER. For example, Hazarika et al. used connected memory networks (ICON) to model the self and inter-speaker influences for participants in a dyadic conversation [7]; Poria et al. proposed a hierarchical RNN framework (DialogRNN) that recurrently models the emotion of the current utterance by considering speaker states and the emotions of preceding utterances [8]; Yeh et al. incorporated contextual information jointly with the current utterance as an attention mechanism (IAAN) [9]. While these works have exhibited the effectiveness of integrating conversation context for emotion recognition, there are some shortcomings. IAAN and ICON only integrated a fixed-length context and ignored the rest of the flow in the dialog history; DialogueRNN did not model the context truly sequentially because it required every utterance to have a consensus emotion label. Furthermore, all of these approaches model the conversational context with the target utterance together in a complex architecture, which limits their rapid extension.

In this work, rather than model the conversational factors via a model architecture, we abstract ERC as two separate modules: the utterance-based recognition engine and a conversation flow decoder (like an acoustic model with a language decoder in ASR). Specifically, we propose an approximate inference algorithm, dialogical emotion decoder (DED), that decodes each utterance into one of the four emotion categories at inference stage. This decoder is built on three core ideas: the emotion which occurs more frequent in dialog history is more likely to show up again; while not all utterances have consensus labels, the posterior distributions capturing affective information would enable us to decode utterances in sequence. Lastly, inspired by [10], the emotion states of interlocutors in a dialog are interleaved and should be jointly modeled. Hence, when given a well-performing SER module, we can then rescore the 4-class probability distribution with proposed DED through a dialog. This method reaches 70.1% UAR on the IEMOCAP [11] (3.0% better than without DED), and 40.3% UAR on the multiparty corpus of the MELD [12]. Since DED is a re-scoring mechanism for ERC, it can also be easily integrated with other variants of SER engines.

| Dataset | Ang | Hap/Joy | Neu | Sad | Ave. speaker per dialog | Ave. dialog length |
|---|---|---|---|---|---|---|
| IEMOCAP | 1103 | 648 | 1708 | 11084 | 2 | 49.2 |
| MELD | 1607 | 2308 | 6436 | 1002 | 2.7 | 9.6 |

**Table 1**. The label distribution and statistics of the IEMOCAP and the MELD.

## 2. METHODOLOGY

### 2.1. Dataset Description

In this paper, we conduct experiments on two different datasets: the IEMOCAP [11] and the MELD [12].

**IEMOCAP** is a benchmark dataset widely used in the field of SER. It contains five sessions with two speakers engaging in different conversational scenarios in each dialog. In this paper, we consider four categories as our classification target: $anger$, $happiness$, $neutral$ and $sadness$. While the results are reported on those utterances with consensus labels, all utterances are used in the decoding process.

**MELD** is a multi-party conversational dataset collected from TV-series, 'Friends'. The database is already split in training, development and testing sets annotated with seven emotions. Here, we consider the same four categories as the IEMOCAP: $anger$, $joy$, $neutral$ and $sadness$.

Table 1 summarizes various key statistics of the two datasets, and we also note that the average dialog length is much shorter in the MELD as compared to the IEMOCAP.

### 2.2. Task Definition

Given a dialog $U = \{u_1, ..., u_T\}$, at each time $t \in [1, T]$ our goal is to recognize $y_t$, i.e., the emotion of $u_t$, depends on the sequence of preceding $t-1$ predicted emotion states $Y_{1:t-1} = \{y_1, ..., y_{t-1}\}$ across $U$. The probability generated for $Y$ is calculated as:

$$p(Y, Z) = p(y_1|x_1) \prod_{t=2}^{T} p(y_t|x_t)\, p(y_t,\, z_t|Y_{1:t-1}, Z_{1:t-1}) \quad (1)$$

where $Z = \{z_1, ..., z_T\}$ with $z_t \in \{0, 1\}$ is a indicator random variable, $z_t = 1$ as speaker's emotion differs from his/her previous one, $z_t = 0$; otherwise. $z_1$ is defined as 1. $x_t$ is the current data point. Additionally, the probability of recognizing the $t^{th}$ emotion state $y_t$ can be factorized into:

$$\begin{aligned} &p\left(y_t|x_t\right) p(y_t,\, z_t|Y_{1:t-1}, Z_{1:t-1}) \\ &\quad = p(y_t|x_t)\, p(y_t|z_t, Y_{1:t-1})\, p(z_t|Z_{1:t-1}). \end{aligned} \quad (2)$$

Here $p(y_t|x_t)$, $p(y_t|z_t, Y_{1:t-1})$, $p(z_t|Z_{1:t-1})$ model the current utterance emotion distribution, probability of emotion assignment based on the past predicted utterances, and the probability of shift (change) in emotion state across time, respectively. Figure 1 shows an overview of emotion decoding process. Each module will be elaborated in Section 2.3.

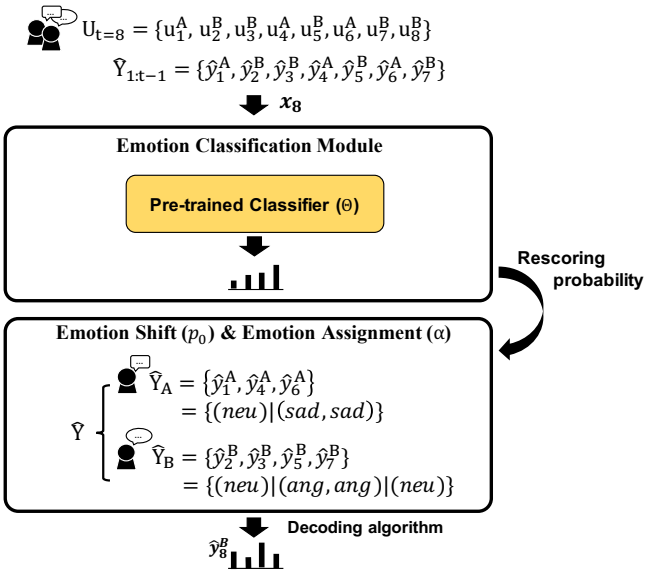


**Fig. 1**. An illustration of dialogical emotion decoder: including both emotion assignment based on prior predicted sequence and emotion shift that models the probabilistic change of emotion states over time.

### 2.3. Dialogical Emotion Decoder (DED)

#### *2.3.1. Emotion Classification Module*

The decoding process of DED is built on top of a SER model. With a pre-trained emotion classifier parametrized by $\Theta$, this classifier is learned to predict a probability distribution of current utterance over four emotion categories given the current behavior sample $x_t$ and $\Theta$:

$$p\left(y_t = l|x_t, \Theta\right), \, l \in E, \quad (3)$$

where $E = \{ang, hap, neu, sad\}$. In this paper, we use IAAN [9] as our emotion classification module. Here $x_t$ is defined as a triple of $(u_c, u_p, u_r)$, the current utterance $u_c$, the previous utterance of the current speaker $u_p$ and the previous utterance of the interlocutor $u_r$. IAAN employs two GRUs for the current utterance and preceding utterances of the speaker and the interlocutor. It leverages the contextual information and affective influences from previous utterances through an attention network that better models the emotion of the current utterance. To lower the computational cost, IAAN is retrained by using unidirectional GRU in modeling the current utterance instead of bidirectional GRU. While our classification module is based on IAAN, DED is not restricted to IAAN; the classification module can be replaced with any other 4-class emotion classifiers.

#### *2.3.2. Emotion Shift Modeling*

While several works have modeled the emotion flow, i.e., often considered emotion follows a smooth transition in conversation [6, 7, 8, 9], these works have not paid attention to

the present of an emotion shift for an individual speaker. For example, in Figure 1, observed from two speakers' utterance sequences $Y_A, Y_B$, there are emotion shifts ($z_t = 1$) at $t = 4$ in $Y_A$ and $t = 3, 7$ in $Y_B$. To account for these changes of states, we model the emotion shift of speakers at time $t$ in a dialog by introducing a bernoulli (binary) distribution with parameter $p_0$:

$$z_t \overset{i.i.d.}{\sim} Bernoulli\,(p_0)\,. \tag{4}$$

We can estimate $p_0$ from the training data:

$$\hat{p}_0 = \frac{\sum_{i=1}^{N}\sum_{p=1}^{M}\sum_{j=2}^{T_{i,p}} \mathbb{1}[y_{i,j}^{p} = y_{i,j-1}^{p}]}{\sum_{i=1}^{N}\sum_{p=1}^{M}(T_{i,p}-1)}, \tag{5}$$

where $N$, $M$ represent the number of training dialogs and the number of participants in that dialog respectively. $T_{i,p}$ is the number of speaker $p$'s utterances in $i^{th}$ dialog. $y_{i,j}^{p}$ is the $j^{th}$ emotion states of speaker $p$ in $i^{th}$ dialog. In addition, in the estimation, $y_{i,j}^{p}$ does not necessary belong to $E$ [11, 12].

#### *2.3.3. Emotion Assignment Process*

Past works only consider short local context in a dialog to improve the SER performance and ignore the rest of the past sequence in the conversation. In this paper, we utilize a distance-dependent Chinese restaurant process (ddCRP) [13] for emotion assignment. It is a clustering approach that has been utilized in image segmentation [14], text modeling [15] and speaker diarization [10]. We use it to model the emotion turns of speakers in a dialog with bounded-interleaved states: $anger$, $happiness$, $neutral$, $sadness$. Regarding those utterances that do not have consensus labels in these four classes, we estimate their probability distribution over the four emotions with pre-trained IAAN. During the assignment process, states can only jump between these four categories, and the occurred state is assigned to corresponding emotion block. The probability assigned to current state $y_t$ is determined as follows:

$$p(y_t = l | z_t = 1, Y_{1:t-1}) \propto \begin{cases} N_{l,t-1}, & l \in Y_{1:t-1} \\ \alpha, & l \in new\,state \end{cases}, \tag{6}$$

where $N_{l,t-1}$ denotes the size of the emotion block for emotion $l$ up to time $t-1$. An emotion block is defined as a sequence that has the longest common-emotion consecutive utterances uttered by an individual speaker.

As illustrated in Figure 1, we have a dialog $U$ of two speakers A and B. In this case, there are two blocks in $Y_A$ and three blocks in $Y_B$, the size of the emotion block for emotion $l$ is defined as $N_{l,t-1} = N_{l,t-1}^{A} + N_{l,t-1}^{B}$. Thus, we have $N_{ang,7} = 0{+}1$, $N_{neu,7} = 1{+}2$, $N_{sad,7} = 1{+}0$, and the probability assigned to a state shown in preceding $t-1$ predicted states $Y_{1:t-1}$ is proportional to its block size. On the other hand, for states not shown in $Y_{1:t-1}$, we assign them a probability that is proportional to a constant $\alpha \in \mathbb{R}$. When $z_t = 0$, speaker's current emotion remains unchanged as his/her previous one. Let $C_t$ be the number of unique present states up to $t$, note that the maximum of $C_t = 4$. The joint distribution of $Y$ under the condition $Z$, $\alpha$ is:

$$p(Y|Z,\alpha) = \frac{\alpha^{C_T-1}\prod_{l\in E}\Gamma(N_{l,T})}{\prod_{t=2}^{T}(\sum_{l\in E}N_{l,t-1}+\alpha)^{\mathbb{1}[z_t=1]}}. \tag{7}$$

Eq.(7) has the same mathematical form as Eq.(8) in [10], but bounded to $E$. In general, we set $\alpha = 1$ for a new state.

#### *2.3.4. Decoding*

Given a testing dialog $U$, we treat $U$ as a sequence and select emotions from all possible states $E$ that maximizes posterior probability from the emotion classification module, emotion shift modeling, and emotion assignment process (Figure 1).

$$\hat{Y} = \arg\max_{Y} \log p(X, Y) \tag{8}$$

Instead of decoding the optimal sequence by exhaustively searching through all the possible outcomes, we adapt greedy search decoder that simply picks the most likely observation at each time $t$ as DED baseline.

$$\begin{aligned}\hat{y}_t, \hat{z}_t = \arg\max_{y_t, z_t} (&\log p(y_t|x_t) + \\ &\log p(y_t|z_t, \hat{Y}_{1:t-1}) + \log p(z_t)).\end{aligned} \tag{9}$$

For better performance, we utilize beam search decoder [16] with beam size $n$; also, before decoding, we duplicate a dialog $K$ times and concatenate them together. We return the prediction of last duplicated sequence as similarly done in [10]. With the concatenation of duplicated dialogs, the probabilities of choosing an existing state would converge and allow better assignment. However, in real-time decoding scenario, $K$ and $n$ should be set to 1.

## 3. EXPERIMENTS AND RESULTS

### 3.1. Experimental Setup

The choice of our 4-class emotion classification module is IAAN as detailed in [9]. We retrain IAAN with the same hyperparameter settings in both datasets but change the Bi-GRU encoder to a GRU. $\alpha$ is set to 1 in emotion assignment process. Greedy search and beam search decoder are utilized along with DED. Also, $K$ is set to 2 in all decoding methods.

The performance is evaluated with unweighted accuracy (UA) and weighted accuracy (WA), where UA represents the average of accuracies of each category, WA represents the percentage of correctly classified samples. For the IEMOCAP, we carry out a 5-fold leave-one-session-out (LOSO) cross validation to evaluate the performance on new conversations with new speakers. Besides, $\hat{p}_0$ in Eq. (5) is estimated from four training sessions. As to the MELD, we adopt early

| Dataset | Method | Recall(%) | | | | WA(%) | UA(%) |
|---|---|---|---|---|---|---|---|
| | | Anger | Happiness | Neutral | Sadness | | |
| IEMOCAP | IAAN | 72.3 | 64.6 | 52.1 | 79.7 | 65.2 | 67.1 |
| | DED$_{Ass}$ + Greedy | 74.8 | 55.5 | 46.0 | 71.7 | 59.6 | 62.0 |
| | DED$_{Shift}$ + Greedy | 70.8 | 66.2 | 54.6 | 80.5 | 66.3 | 68.0 |
| | DED + Greedy | 68.7 | 68.8 | 59.5 | 79.5 | 68.0 | 69.1 |
| | DED + Beam Search ($n = 10$) | 70.3 | 70.5 | 60.0 | 79.6 | **69.0** | **70.1** |
| MELD | IAAN | 43.8 | 30.8 | 43.4 | 39.8 | 40.8 | 39.4 |
| | DED$_{Ass}$ + Greedy | 35.5 | 27.8 | 45.9 | 35.1 | 39.8 | 36.9 |
| | DED$_{Shift}$ + Greedy | 41.6 | 29.9 | 41.3 | 38.9 | 39.1 | 37.9 |
| | DED + Greedy | 36.9 | 34.5 | 43.7 | 38.8 | 40.1 | 38.4 |
| | DED + Beam Search ($n = 20$) | 40.8 | 29.7 | 49.4 | 41.4 | **43.6** | **40.3** |

**Table 2**. The performance of using classification module of IAAN with DED using greedy and beam search decoder. We evaluate the performance on the IEMOCAP with LOSO cross validation, and on the pre-defined testing set for the MELD.

stopping criterion by observing the performance on the validation set every $100$ training steps, then we predict on the pre-defined testing set. We estimate $\hat{p}_0$ from training set and conduct DED on testing set. Finally, we evaluate the decoded utterances labeled in the desired emotion set $E$.

### 3.2. Comparison Methods

To examine the effectiveness of different components of DED, we further present the results of its variants, i.e., decoder with emotion shift only or with emotion assignment only, denoted by DED$_{Shift}$ and DED$_{Ass}$ respectively. We experimented DED and its variants with two approximate search algorithms: greedy search and beam search decoders. IAAN is used as our vanilla baseline, which includes only short conversational context information. We set beam size $n$ to 10 in the IEMOCAP. In the MELD, since the performance of IAAN is lower than in the IEMOCAP, we use a larger beam size, $n = 20$. Table 2 summarizes the results of different approaches measured in UA and WA.

### 3.3. Result and Analysis

From Table 2, DED+Beam Search performs better than baseline IAAN on both datasets. Specifically, in the IEMOCAP, DED+Beam Search reaches $69.0\%$ in WA and $70.1\%$ in UA with relative $3.8\%$, $3.0\%$ improvement over vanilla IAAN. Furthermore, with a simple greedy search decoder, DED+Greedy performs $69.1\%$, which obtains an absolute $2\%$ improvement over IAAN in UA measure. In the MELD, DED+Beam Search reaches $43.6\%$ in WA and $40.3\%$ in UA, obtaining a relative $2.8\%$, $0.9\%$ improvement over IAAN. On the other hand, greedy approach shows no improvement, and falls by $1\%$ in UA. By examining different DED variants, we observe that DED$_{Shift}$ performs $6\%$, and $1\%$ higher UA than DED$_{Ass}$ in the IEMOCAP and the MELD respectively. The proposed DED, i.e., incorporating the assignment process conditioned on a binary indicator $z_t$ as Eq (7) shown, obtains $1.1\%$, $0.5\%$ higher UA than DED$_{Shift}$ in both datasets.

The comparison between DED$_{Shift}$ and IAAN indicates that by considering the previous predicted emotion state of an individual speaker, a bernoulli distribution can effectively embed the transition information to the 4-class probability distribution. Despite the lower performance observed in DED$_{Ass}$ compared with DED$_{Shift}$, the combination of two modules can further improve the performance of DED$_{Shift}$, especially in neutral class. The neutral category can be hardly modeled by IAAN due to lack of apparent emotional characteristics [9, 17, 18]. With the help of emotion shift combined with emotion assignment process, DED is able to rescore the posterior distributions through a dialog based on previous predicted emotion states. However, as reported in Table 2, the effectiveness of DED in the MELD is not as significant as in the IEMOCAP. First, DED relies on a well-performing classifier, due to the poorer performance of IAAN in the MELD (likely caused by the noisy recordings of TV series that include other sound effect), the effect of DED is more limited. Furthermore, as the MELD database is collected from much shorter sequence of dialogs (average length is less than 10 utterances), the ability of DED in properly decoding through long sequence of dialogs is not as obvious as for the IEMOCAP database (average length is about 50 utterances).

## 4. CONCLUSION

In this paper, we propose a dialogical emotion decoding algorithm that performs emotion decoding in a dialogical manner. Compared with other methods in handling ERC, we can model the emotion flows in a dialog consecutively by combining emotion classification, emotion shift and emotion assignment process together. Our method exhibits outstanding performance on four emotion class UA of $70.1\%$ in the IEMOCAP. In addition, we evaluate DED on a multi-party dataset MELD, and our proposed DED demonstrates improvement over baseline classification method. Our proposed DED is an inference algorithm with its classification module replaceable with other pre-trained utterance based emotion recognition engines. Our future research directions include investigating the robustness of DED on different emotion engines, modeling the emotion shift with model-based approaches instead of a bernoulli distribution.

## 5. REFERENCES


[1] Susan Shott, "Emotion and social life: A symbolic interactionist analysis," *American journal of Sociology*, vol. 84, no. 6, pp. 1317–1334, 1979.

[2] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2227–2231.

[3] Michael Neumann and Ngoc Thang Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *arXiv preprint arXiv:1706.00612*, 2017.

[4] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5200–5204.

[5] Jonathan Chang and Stefan Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2746–2750.

[6] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *arXiv preprint arXiv:1905.02947*, 2019.

[7] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann, "Icon: Interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2594–2604.

[8] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 6818–6825.

[9] Sung-Lin Yeh, Yun-Shao Lin, and Chi-Chun Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6685–6689.

[10] Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang, "Fully supervised speaker diarization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6301–6305.

[11] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[12] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[13] David M Blei and Peter I Frazier, "Distance dependent chinese restaurant processes," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2461–2488, 2011.

[14] Soumya Ghosh, Andrei B Ungureanu, Erik B Sudderth, and David M Blei, "Spatial distance dependent chinese restaurant processes for image segmentation," in *Advances in Neural Information Processing Systems*, 2011, pp. 1476–1484.

[15] Georgios Palaiokrassas, Athanasios Voulodimos, Antonios Litke, Athanasios Papaoikonomou, and Theodora Varvarigou, "A distance-dependent chinese restaurant process based method for event detection on social media," *Inventions*, vol. 3, no. 4, pp. 80, 2018.

[16] Mark F. Medress, Franklin S Cooper, Jim W. Forgie, CC Green, Dennis H. Klatt, Michael H. O'Malley, Edward P Neuburg, Allen Newell, DR Reddy, B Ritea, et al., "Speech understanding systems: Report of a steering committee," *Artificial Intelligence*, vol. 9, no. 3, pp. 307–316, 1977.

[17] Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010.

[18] Anne-Maria Laukkanen, Erkki Vilkman, Paavo Alku, and Hanna Oksanen, "On the perception of emotions in speech: the role of voice quality," *Logopedics Phoniatrics Vocology*, vol. 22, no. 4, pp. 157–168, 1997.