Audio and ASR-based Filled Pause Detection

Aggelina Chatziagapi
Stony Brook University & Behavioral Signal Technologies
New York, U.S.A
aggelina@cs.stonybrook.edu

Dimitris Sgouropoulos
Behavioral Signal Technologies
Athens, Greece
dimitris@behavioralsignals.com

Constantinos Karouzos
Behavioral Signal Technologies
Athens, Greece
constantinos@behavioralsignals.com

Thomas Melistas
Behavioral Signal Technologies
Athens, Greece
thomas@behavioralsignals.com

Theodoros Giannakopoulos
Behavioral Signal Technologies & NCSR Demokritos
Athens, Greece
thodoris@behavioralsignals.com

Athanasios Katsamanis
Behavioral Signal Technologies
Athens, Greece
nassos@behavioralsignals.com

Shrikanth Narayanan
Behavioral Signal Technologies
Los Angeles, U.S.A
shri@behavioralsignals.com

2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII) | 978-1-6654-5908-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ACII55700.2022.9953889

***Abstract*—Filled pauses (or fillers) are the most common form of speech disfluencies and they can be recognized as hesitation markers ("um", "uh" and "er") made by speakers, usually to gain extra time while thinking their next words. Filled pauses are very frequent in spontaneous speech. Their detection is therefore rather important for two basic reasons: (a) their existence influences the performance of individual components, like Automatic Speech Recognition system (ASR), in human-machine interaction and (b) their frequency can characterize the overall speech quality of a particular speaker, as it can be strongly associated with the speaker's confidence. Despite that, only limited work has been published for the detection of filled pauses in speech, especially through audio. In this work, we propose a framework for filled pause detection using both audio and textual information. For the audio modality, we transfer knowledge from a plethora of supervised tasks, such as emotion or speaking rate, using Convolutional Neural Networks (CNNs). For the text modality, we develop a temporal Recurrent Neural Network (RNN) method that takes into account textual information derived from an ASR system. In addition, the proposed transfer learning approach for the audio classifier leads to better results when benchmarked on our internal dataset for which the text is not transcribed but estimated by an ASR system. In this case, a simple late fusion approach boosts the performance even further. This proves that the audio approach is suitable for real-world applications where the transcribed text is not available and has to leverage imperfect ASR results, or even the absence of textual information (to reduce computational cost).**

***Index Terms*—Filled Pauses, Hesitation, Disfluency Detection, Audio Classification, Text Classification, Multimodal Learning, Automatic Speech Recognition, Convolutional Neural Networks, Recurrent Neural Networks, Deep Learning**

I. Introduction

With the latest advances in Machine Learning, Natural Language Processing (NLP), and Automatic Speech Recognition (ASR), modern speech analysis systems can recognize what people say while also analyzing the semantics. Special focus has recently been given on understanding *how* people talk to machines and *how* people interact with each other. To extract such non-verbal information, systems take into account prosodic cues and attempt to detect speech attributes that are independent of language and culture and more related to behaviors and emotions. The goal is to enhance the overall human-machine communication, as well as the performance of individual components, like ASR.

A common characteristic of conversational speech that affects the human-machine interaction is the presence of disfluencies. Disfluencies are prosodic and grammatical interruptions and include silent pauses, repetitions ("I was I was"), repairs ("I was we were"), and restarts ("I was Today is...") [1]. In addition, they may involve hesitation markers, such as "um", "uh" and "er", which are also referred to as *filled pauses* or *fillers* ("I was um I was"). Hesitation markers are very common in Germanic languages and they have long been studied in linguistics [2]. Their occurrence is usually related to the cognitive processes responsible for the production of speech. For example, the speaker may use such a marker for speech planning or search of better wording. Furthermore, recent studies have proved that filled pauses might be beneficial for listeners to comprehend the meaning of a spoken utterance faster [2], [3]. Apart from the communicative functions, disfluencies can also express the speaker's mental state [4].

Because of their frequency in spontaneous speech, disfluencies pose a number of challenges when it comes to real-life human-machine communication. First, they are not usually handled adequately by the language models of ASR systems. Typical speech recognition systems are trained to deal with fluent speech, without considering hesitation phenomena [4]. Second, their presence in synthetic speech is essential, in order to reproduce the spontaneity of natural human interaction [5].

However, it is not evident for speech synthesizers where to insert filled pauses or how to modify the prosody to adjust the fluency of the output speech. An inherent challenge, that affects both speech recognition and synthesis, concerns the annotation of disfluencies. Broadly speaking, their occurrence can be subjective depending on their duration and they can be difficult to notice [6]. Thus, the available corpora annotated for relevant tasks are rather limited. Recent studies have proposed semi-automatic annotation as a solution to improve the generation of such datasets [6], [7].

Developing classifiers for disfluency detection can be significantly beneficial towards addressing the aforementioned issues. Nevertheless, the related work in the literature is quite limited, especially for audio-based classification. The authors of [4] propose a real-time filled pause detection system using only two acoustic features, namely fundamental frequency and spectral envelope deformation. The work of [8] integrates masking and filtering techniques on probabilistic time series to detect filler and laughter events from phone calls. Filler and laughter detection have gained further attention as a subtask of social signal detection in [9]. The researchers in [10] utilized a Bidirectional Long Short-Term Memory (BLSTM) network with Connectionist Temporal Classification (CTC) loss to detect social signals without the need of alignment labeling, while in [11] the performance of BLSTMs and Gated Recurrent Units (GRUs) is explored in the mono-lingual and cross-lingual setting.

Regarding text-based approaches, [12] introduces a BLSTM network for disfluency detection based on the Switchboard corpus [13]. Similarly, [14] uses an incremental transition-based framework. More recent works, such as [15], incorporate self-training and ensembling, utilizing a pre-trained BERT [16]. Moreover, some multimodal methods have also been proposed. For example [1] combines word embeddings of trascribed text and acoustic-prosodic features, including pause and word duration. Video information is used by [17] along with audio for filler detection in classroom lectures.

In this paper, we propose an end-to-end solution for filled pause detection based on both audio and text information. However, we do not assume that the textual information is available as a zero-error transcription. Instead, we face the problem of text-based filled pause detection in the context of a real-world scenario, according to which *text is estimated* by an ASR system and it is therefore prone to errors. We present experimental results for separate audio and text-based detection, using deep audio and text classifiers correspondingly, as well as for a late fusion approach that combines their outputs. The main contributions of the proposed approach are the following:

1) A deep learning model based on *audio* data is presented. Results on both open and proprietary datasets prove that this method outperforms the text classifier when text is derived from an ASR system, i.e. in a realistic case of speech analysis.
2) A new approach for text classification of ASR output is proposed, based on a temporal representation and an RNN architecture.

II. Proposed Methodology

A. Overall Architecture

Figure 1 shows a conceptual diagram of the proposed end-to-end filled pause detection pipeline. For each speech segment, our objective is to detect the existence of filled pauses from the raw audio data. With regards to the audio modality, this is achieved through direct audio analysis, extracting a spectrogram and training a Convolutional Neural Network. In addition, the audio information is fed as input to an ASR module that predicts spoken words and respective time stamps (speech-to-text). The sequence of (words, time stamps) pairs is used by a text classifier that similarly predicts if the respective speech segment contains fillers or not. The final decision is taken by a simple late fusion layer.

The text classifier can be any typical classifier. However, in this work, we propose a new approach for the text representation that also takes into account the respective time stamps of each recognized word. The proposed pipeline makes no assumption about the existence of speech transcriptions (i.e. correct and human-transcribed speech-to-text). This is the focus of this work with regards to the experimental setup as well: apart from evaluating both audio and text classifiers on an open dataset for the sake of comparison, we have also conducted detailed experimentation on a diverse-domain, real-world dataset where human transcriptions are unavailable. The audio and ASR-based decisions are then fused using a simple posterior averaging step.

B. Text-based Detection

1) Word RNN: Recurrent Neural Networks (RNNs), especially Long Short-Term Memory units (LSTMs), have been widely used for representing sequential data in various fields, ranging from emotion classification [18], [19] to machine translation [20]. Their ability to capture temporal information has been proven truly significant. In this work, we adopt LSTMs for the text classification task, to detect disfluency in sequences of words.

More specifically, we represent each utterance with pre-trained word embeddings, namely GloVe embeddings [21]. The resulted representation is passed to a Bidirectional LSTM (BLSTM) followed by an attention module and a feedforward network. The attention module is based on the architecture proposed by [20]. It consists of 2 linear layers with a hyperbolic tangent (tanh) activation function in between and a softmax activation with a single output. In this way, each input of the feedforward network is calculated as the weighted sum of the BLSTM's hidden states over all time sequences. The feedforward network includes 3 linear layers. After the first 2 linear layers, we apply ReLU activation. The last linear layer is followed by a softmax activation function, which predicts if the input utterance contains filled pauses or not. Dropout with $p = 0.2$ is applied after all the intermediate layers of the whole model. Adam optimizer with a learning rate of 10^{-3} is used for training.

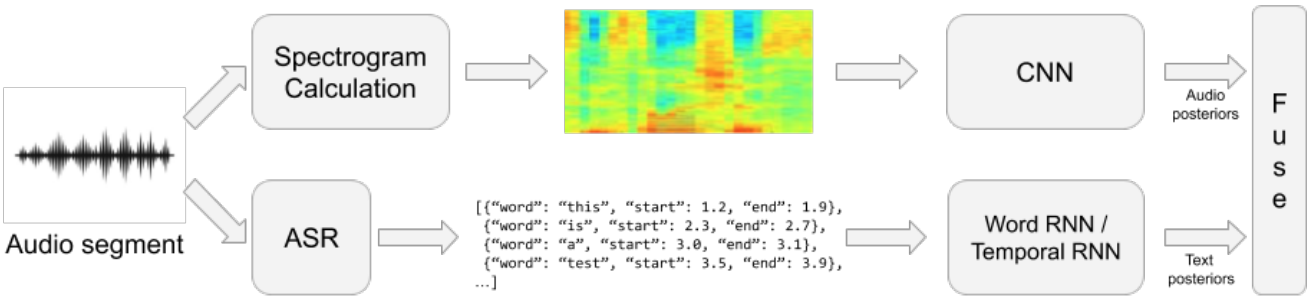

Fig. 1. Conceptual diagram of the proposed audio and ASR-based filled pause classification approach

2) Temporal RNN: In addition to the word embeddings, we propose a temporal representation of the utterances. This representation aims to enhance the input information for the filled pause detection task, by capturing speech and silence on a temporal level. To this end, we split each utterance to fixed-size time windows with a step of 50 msec. We fill each window with the token of the word that occurs in the corresponding time interval. If no word occurs, we fill it with the silence token. To illustrate that, consider a sequence of tokens $[t_1, t_2]$ resulting from the embedding layer for a specific utterance, with time stamps [0 - 200, 270 - 400] in msec. Then, we can arrange the tokens along a time axis as follows:

0-50	50-100	100-150	150-200	200-250	250-300	300-350	350-400
t_1	t_1	t_1	t_1	s	t_2	t_2	t_2

where s is the silence token, the first row corresponds to the time stamps in msec and the second row corresponds to the temporal representation of this utterance.

Using this representation, we train a text classifier that follows the same architecture as described in Sec. II-B1, consisting of a BLSTM, an attention module and a feedforward network. We call this classifier *temporal RNN*. Furthermore, to leverage the information of both the GloVe and temporal representations, we construct an additional text classifier, namely *word-temporal RNN*, that includes separate BLSTMs and attention modules for each encoding, concatenates the outputs and feeds a shared feedforward network. The purpose is to leverage any synergy between the two different types of text representations.

C. Audio-based Detection

1) CNN for Segment Classification: During the last years, Convolutional Neural Networks (CNNs) have proven their ability to represent robust and invariant image features, and due to that, they have become the most widely adopted classification and supervised feature extraction technique in computer vision [22]. Their ability to capture features from multi-dimensional spaces, has recently led to the adoption of CNNs into speech and audio classification tasks [23]–[25], as well as music analysis [24] and emotion recognition [26]–[28]

In this work, we adopt CNNs as classifiers of fixed-size audio segments. Each audio segment of 3 seconds duration is first represented as a mel-scaled spectrogram. For the spectrogram extraction, we use a short-term window of 50 msec with a 50% overlap ratio, while the number of Mel coefficients is 128. This results in fixed-size spectrograms of 128×121. After the frequency power calculation, we apply logarithmic scale and z-normalization based on the statistics of the training data.

Regarding the CNN architecture, it consists of 4 convolutional and 2 linear layers, as demonstrated in Fig. 2. After each convolutional layer, we add batch normalization, ReLU activation, max pooling, and dropout with $p = 0.2$. After the first linear layer, we similarly apply ReLU activation and dropout. The last linear layer is followed by a softmax activation function, which attempts to predict if the input segments contain disfluency or not.

2) Transfer Learning: As with most real-world supervised classification tasks, it is rather expensive to collect and annotate sufficiently large and diverse corpora of speech samples for filled pause classification. In such cases, transferring knowledge from a similar task, a process known as "transfer learning", is the most straightforward solution. The source task usually has a large amount of labeled data, which enable to train a model that performs well in this task. The target task has significantly less data, rendering the training from scratch of a deep neural network difficult. Transfer learning has been widely used in the context of CNNs and other deep learning architectures in various domains, including computer vision and speech analysis [25], [29].

In this work, we conduct experiments transferring knowledge to the filled pause detection task from other speech analysis tasks. More specifically, we train classification models for (a) speaking rate, (b) emotion, (c) valence, and (d) arousal, as well as (e) voice activity detection (VAD), using approximately $160k$ speech utterances of over 300 hours total duration. Then, we initialize the model for our target task with the learned weights and fine-tune the layers' parameters on the filled pause dataset. For both the initial training in the source domain and the fine-tuning in the target domain, we use Adam optimizer with a learning rate of 10^{-4} and a weight decay of 10^{-6}.

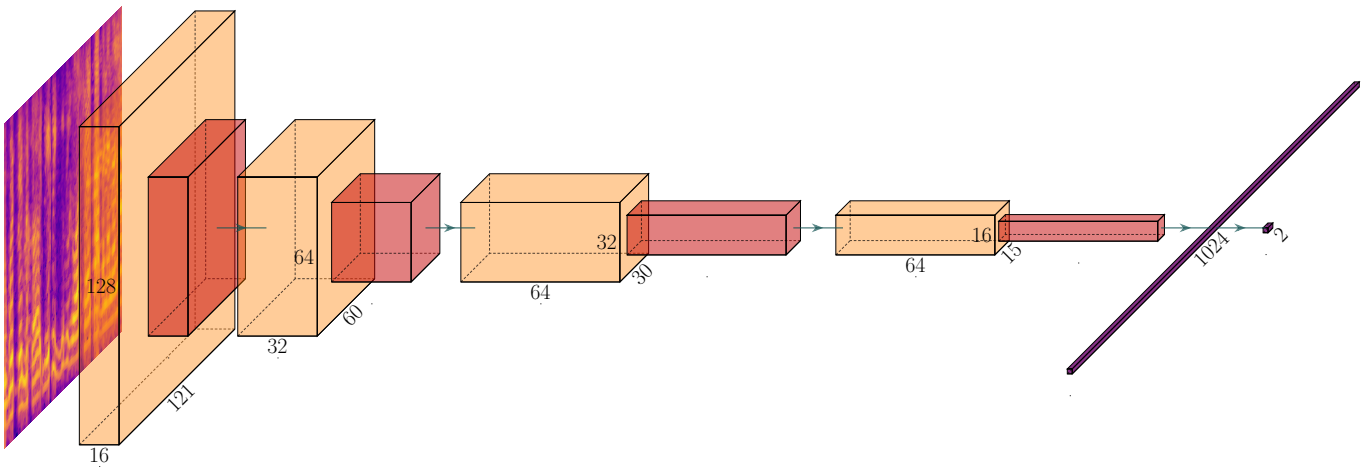

Fig. 2. CNN architecture for audio-based filled pause detection

D. Class-Balanced Loss

Real-world data often suffer from class imbalance, so to alleviate that problem during model training we apply weights to the cross-entropy loss function. In this way, we encourage the network to focus more on the underrepresented classes, which have significantly less samples than the majority class. Typically, a class-balanced loss function assigns class weights that are inversely proportional to the class populations [30]. Following this strategy, we have noticed that the models tend to be biased towards recall. To avoid that, we adopt a smoothed class weighting scheme. Specifically, for each class c with n_c training samples we assign a weight as follows:

$$w_c = \frac{\sum_{i=1}^{C} n_i^{\alpha}}{n_c^{\alpha}} \tag{1}$$

where C is the total number of classes and α is a smoothing factor in the range $[0, 1]$. For the datasets in this work, $\alpha = 0.25$ led to the best performance with a balance between recall and precision.

E. Aggregation and Fusion

Having trained an audio and text classifier for filled pause detection, we proceed with comparing their performance on an utterance level. Depending on the dataset, an utterance may be longer than 3 seconds. In such cases, we extract the segment-level predictions of the audio classifier, after breaking each utterance to 3-second segments with a 2-second overlap. We aggregate these predictions by applying majority voting.

In addition to the audio and text-based classification, we also consider their fusion. In the last years, various fusion techniques have been proposed in the literature [1], [31], [32]. In this work, we follow a simple late fusion approach, combining the posteriors of the audio and text models. Given an utterance with p_a and p_t the posteriors of the audio and text classifiers correspondingly, we calculate their weighted average:

$$score = w \cdot p_a + (1 - w) \cdot p_t \tag{2}$$

where w is a scalar in the range $[0, 1]$. The utterance is classified to the class with the maximum score. Regarding w, we try values from 0.1 to 0.9 with a step of 0.1 and select the one that results in the maximum F-score for the validation set.

III. Experiments

A. Datasets

1) Internal Datasets: The following internal and proprietary datasets have been used for training the filled pause audio and text classifiers, as well as audio classifiers used for transfer learning to the target classification task of filled pause detection. Note that in all cases, apart from the raw audio signal, text information is available in the form of ASR predictions (not human transcriptions), which has been extracted by our internal ASR system. This system has a word error rate (WER) of around 40% on the particular real-world data, ranging from 10% up to 60% depending on the domain and its recording and contextual conditions.

i) *Filled pauses*: This is the main dataset of this paper. It contains approximately $38k$ audio samples in English, gathered from real-life conversations. Each audio sample corresponds to a 3-second segment that is annotated for occurrence of disfluency. The audio sample duration was chosen to make the annotation procedure easier and more robust. We mainly consider filled pauses for this dataset, but there are also cases of other types of hesitation, such as word lengthenings, repetitions and repairs. Each segment has been labeled calculating the majority vote of 3 to 9 human annotators. We split the filled pauses dataset with a ratio of 60% - 40% for training and testing correspondingly in a speaker-disjunct way.

ii) *Speaking rate*: The speaking rate dataset consists of exactly the same audio samples as the filled pauses dataset.

But, in this case, each segment is annotated in terms of speaking rate, considering the following classes: $slow$, $normal$, and $fast$. Even though this dataset is relatively small, we used it as the source task for transfer learning, because of its similarities with the filled pause detection. Specifically, segments annotated as slow have a high probability to contain hesitations or pauses. In contrast, the ones annotated as fast may contain more confident speech, without any disfluencies. In addition, it produced a stable model for speaking rate classification, with a performance of almost 70% on the test set.

iii) *Emotion*: The emotion dataset contains speech utterances from spontaneous real-life conversations gathered from several domains. Each utterance has a duration of 3 to 10 seconds, with an average of 7 seconds, and it is annotated in terms of the speaker's emotional state. We consider the following 5 emotional classes: $angry$, $happy$, $neutral$, sad, and $ambiguous$. The latter corresponds to samples for which the inter-annotator agreement was lower than a particular threshold. Each sample has been labeled by 3 to 7 human annotators. Overall, the dataset includes more than $160k$ utterances of over 300 hours total duration. Due to its real-world nature, a class imbalance between the neutral and emotional classes is expected. More specifically, the ratio of the less populated (sad) to the most dominant (neutral) class is 3%.

iv) *Valence*: The valence dataset is composed of exactly the same audio data as the emotion dataset. They are annotated in terms of valence considering the following classes: $negative$, $neutral$, $positive$, and $ambiguous$. Similarly to emotion, samples are labeled as $ambiguous$ if the inter-annotator agreement was lower than a threshold. The ratio of the minority (positive) to the majority (neutral) class is 4%.

v) *Arousal*: Similarly, the arousal dataset contains the same speech utterances as the emotion and valence datasets, but they are annotated for arousal. The arousal classes are: $weak$, $neutral$, $strong$, and $ambiguous$. The ratio of the minority (weak) to the majority (neutral) class is 4%.

vi) *VAD*: The VAD dataset consists of speech utterances gathered from the same broad domains as the emotion dataset and follows a similar structure. Each chunk is labeled on the basis of containing or not containing speech. The non-speech class includes silence, music, and noise. Depending on the domain and the recording conditions, there are various types of noise that occur in the dataset. In total, it contains approximately $170k$ utterances with a duration of over 300 hours.

2) Open Dataset: In order to compare the proposed method with other works in the literature, we additionally conduct experiments using the widely used Switchboard corpus [13]. Switchboard consists of spontaneous two-sided telephone conversations among speakers from the US. It includes annotation for various linguistic attributes and it is widely used for speech recognition and other speech analysis tasks. In this work, we use the Switchboard-NXT release [33], which is a subset of the original corpus organized within a unified framework in XML format. It includes 642 conversations that are segmented to approximately $108k$ utterances. Each utterance is annotated for disfluency, which is marked by a reparandum, i.e. the words where the speaker hesitated or made a false start, and a repair, where the speaker corrected the error. For our experiments, we follow the standard data split to train/dev/test sets, as defined by [1]. Note that, the Switchboard corpus includes text transcriptions, not ASR predictions.

B. Performance Results and Discussion

In this section, we present the performance results for both the internal and open datasets. In Table I we present the average F-score achieved on the test set of our internal dataset for filled pause classification. The results of all the audio-based methods are listed, including the simple CNN training and transfer learning from each available dataset. Transferring knowledge from the speaking rate task leads to an improvement of almost 1%, indicating the impact of transfer learning in this setting. In contrast to Switchboard, the text of this dataset has not been produced by human transcription, but by our ASR system and as a result, it can include errors due to the nature of real-wold, noisy conversations. Consequently, it is not surprising that text results are slightly lower than the audio results. As far as fusion is concerned, we combine each text classifier with the best audio classifier, which is the one trained transferring knowledge from the speaking rate task. Every fusion combination leads to a boost in performance, suggesting again that both audio and text information is beneficial. The fusion of the word-temporal RNN and audio classifier demonstrates the best F-score, increasing the performance by 1.4% and 2.8% compared to using only the audio and text modality correspondingly. Last but not least, we evaluated the precision of the latter in terms of the annotation confidence. Specifically, 37% of the incorrectly predicted filled pauses (false positives) were borderline decisions, with an annotation confidence (i.e. the inter-annotator agreement measured during the annotation process) of less than 70%.

TABLE I
PERFORMANCE ON OUR DATASET (F-SCORE %)

Modality	**Method**	**F-score**
6*Audio	CNN	71.6
	Transfer from speaking rate	72.5
	Transfer from emotion	72.2
	Transfer from valence	71.6
	Transfer from arousal	71.9
	Transfer from VAD	71.8
3*Text	Word RNN	67.3
	Temporal RNN	70.2
	Word - Temporal RNN	71.1
3*Fusion	Audio & Word RNN	73.2
	Audio & Temporal RNN	73.6
	Audio & Word - Temporal RNN	**73.9**

In Table II we demonstrate the disfluency detection F-score achieved on the test set of Switchboard. We use this metric to

be comparable with other works in the literature. Regarding the audio classifier, we report the result of the transfer learning from the speaking rate task, since it was the best among all other tasks, as also shown in Table I for our internal dataset. As for the text classifier, the combination of word embeddings and temporal representation resulted in a significant performance improvement of over 4%, compared to using only individual encodings. In addition, the 82.3% detection F-score shows an increase of 4.1% from related work [1], [34]. The fusion results demonstrate a small performance boost when the text classifier involves only one encoding. This suggests that both audio and text information can be beneficial for the disfluency detection task. However, for Switchboard, the word-temporal RNN outperforms its fusion with the audio classifier, which can be justified by the nature of this dataset: it includes text derived from detailed transcriptions, leading to much higher results for text than audio classifiers.

TABLE II
PERFORMANCE ON SWITCHBOARD (DISFLUENCY DETECTION F-SCORE %)

Modality	**Method**	**F-score**
1*Audio	Transfer from speaking rate	68.5
3*Text	Word RNN	78.0
	Temporal RNN	77.7
	Word - Temporal RNN	**82.3**
3*Fusion	Audio & Word RNN	78.9
	Audio & Temporal RNN	79.3
	Audio & Word - Temporal RNN	82.1

IV. CONCLUSIONS AND FUTURE WORK

In this work, we present a deep learning framework for filled pause detection in speech, based on both audio and ASR-generated text information. More specifically, for the audio classifier, we adopt a CNN architecture that is applied on spectrograms and we transfer knowledge from other speech analysis tasks. For the text classifier, we propose a new approach, namely "word-temporal RNN", that leverages temporal information (time stamps) along with word embeddings. We focus on ASR-generated text, which is not a golden ground truth, like the one generated by human transcription. The challenge here is that ASR systems have a non-zero word error rate that leads to a noisy text output, especially in real-world situations where conversations can be spontaneous and recording conditions can vary. Experimentation has taken place on both an open dataset, which includes transcribed text information, and an internal multi-domain real-world dataset, where text was estimated by an ASR system. Results prove that:

1) The word-temporal RNN outperforms previous similar text classifiers from related work in the task of filled pause detection, when evaluated on the open dataset (transcribed text case).
2) The proposed audio classifier has a significantly lower performance on the open dataset, compared to the text, which can be expected considering the fact that the textual information is transcribed in this case.
3) The proposed audio classifier is slightly better in the case of the internal dataset, compared to the best text classifier, which, again is the word-temporal RNN method.
4) The simple late fusion of the audio and ASR-based classifiers leads to 1.4% absolute performance improvement, comparing to the best standalone modality (the audio classifier), for the internal dataset experimentation.

In other words, the proposed framework guarantees accurate filled pause detection and robustness to ASR-related errors, which are frequent in real-world scenarios. Furthermore, we have demonstrated that the audio classifier outperforms the ASR-based approach and is just slightly worse than the fusion estimate. Thus, it can be used to detect filled pauses without the use of ASR, significantly reducing the overall computational cost of the whole processing pipeline.

In the future, we would like to explore the use of the Transformer architecture [35] and specifically pre-trained models on text, such as BERT [16] for the text-based detection component. Finally, for the audio-based detection part, we can leverage the representation already learned by the ASR acoustic model, using it to initialize the filler detector, as well as incorporating attention mechanisms in our classifier to more specifically locate the temporal position of fillers in each utterance.

REFERENCES

[1] T. Tran, S. Toshniwal, M. Bansal, K. Gimpel, K. Livescu, and M. Ostendorf, "Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 69–81.

[2] M. Wieling, J. Grieve, G. Bouma, J. Fruehwald, J. Coleman, and M. Liberman, "Variation and change in the use of hesitation markers in Germanic languages," *Language Dynamics and Change*, vol. 6, no. 2, pp. 199–234, Nov. 2016.

[3] M. Corley, L. J. MacGregor, and D. I. Donaldson, "It's the way that you, er, say it: Hesitations in speech affect language comprehension," *Cognition*, vol. 105, no. 3, pp. 658–668, 2007.

[4] M. Goto, K. Itou, and S. Hayamizu, "A real-time filled pause detection system for spontaneous speech recognition," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[5] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis," in *The 10th ISCA Speech Synthesis Workshop*, 2019.

[6] S. Betz, J. Voße, S. Zarrieß, and P. Wagner, "Increasing recall of lengthening detection via semi-automatic classification," in *Proceedings of Interspeech*, 2017.

[7] O. Egorow, A. Lotz, I. Siegert, R. Bock, J. Krüger, and A. Wendemuth, "Accelerating manual annotation of filled pauses by automatic preselection," in *2017 international conference on companion technology (icct)*. IEEE, 2017, pp. 1–6.

[8] R. Gupta, K. Audhkhasi, S. Lee, and S. S. Narayanan, "Paralinguistic event detection from speech using probabilistic time-series smoothing and masking." 2013.

[9] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[10] H. Inaguma, K. Inoue, M. Mimura, and T. Kawahara, "Social signal detection in spontaneous dialogue using bidirectional lstm-ctc." 2017.

[11] R. Brueckner, M. Schmitt, M. Pantic, and B. Schuller, "Spotting social signals in conversational speech over ip: A deep learning perspective," *Proc. Interspeech 2017*, pp. 2371–2375, 2017.

[12] V. Zayats, M. Ostendorf, and H. Hajishirzi, "Disfluency detection using a bidirectional lstm," *Interspeech 2016*, pp. 2523–2527, 2016.

[13] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1. IEEE Computer Society, 1992, pp. 517–520.

[14] S. Wang, W. Che, Y. Zhang, M. Zhang, and T. Liu, "Transition-based disfluency detection using LSTMs," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017, pp. 2785–2794.

[15] P. J. Lou and M. Johnson, "Improving disfluency detection by self-training a self-attentive model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3754–3763.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[17] V. Tsiaras, C. Panagiotakis, and Y. Stylianou, "Video and audio based detection of filled hesitation pauses in classroom lectures," in *2009 17th European Signal Processing Conference*. IEEE, 2009, pp. 834–838.

[18] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 606–615.

[19] V. Mitra, S. Booker, E. Marchi, D. S. Farrar, U. D. Peitz, B. Cheng, E. Teves, A. Mehta, and D. Naik, "Leveraging Acoustic Cues and Paralinguistic Embeddings to Detect Expression from Voice," in *Proc. Interspeech 2019*, 2019, pp. 1651–1655.

[20] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[21] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[23] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[24] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2392–2396.

[25] M. Papakostas and T. Giannakopoulos, "Speech-music discrimination using deep visual feature extractors," *Expert Systems with Applications*, vol. 114, pp. 334–344, 2018.

[26] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2017.

[27] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, and S. Narayanan, "Data Augmentation using GANs for Speech Emotion Recognition," *Proc. Interspeech 2019*, pp. 171–175, 2019.

[28] J. Lee, J. Park, K. L. Kim, and J. Nam, "SampleCNN: End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification," *Applied Sciences*, vol. 8, no. 1, p. 150, 2018.

[29] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[30] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9268–9277.

[31] J. Sebastian and P. Pierucci, "Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts," in *Proc. Interspeech 2019*, 2019, pp. 51–55.

[32] E. Georgiou, C. Papaioannou, and A. Potamianos, "Deep Hierarchical Fusion with Application in Sentiment Analysis," in *Proc. Interspeech 2019*, 2019, pp. 1646–1650.

[33] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver, "The NXT-Format Switchboard Corpus: A Rich Resource for Investigating the Syntax, Semantics, Pragmatics and Prosody of Dialogue," *Language Resources and Evaluation*, vol. 44, no. 4, p. 387–419, 2010.

[34] J. G. Kahn, M. Lease, E. Charniak, M. Johnson, and M. Ostendorf, "Effective use of prosody in parsing conversational speech," in *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 2005, pp. 233–240.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.