RESEARCH

Language Agnostic Missing Subtitle Detection

Honey Gupta* and Mayank Sharma

Abstract

Subtitles are a crucial component of Digital Entertainment Content (DEC such as movies and TV shows) localization. With ever increasing catalog ($\approx$ 2M titles) and localization expansion (30+ languages), automated subtitle quality checks becomes paramount. Being a manual creation process, subtitles can have errors such as missing transcriptions, out-of-sync subtitle blocks with the audio and incorrect translations. Such erroneous subtitles result in an unpleasant viewing experience and impact the viewership. Moreover, manual correction is laborious, highly costly and requires expertise of audio and subtitle languages. A typical subtitle correction process consists of 1) linear watch of the movie, 2) identification of time stamps associated with erroneous subtitle blocks and 3) correcting procedure. Among the three, time taken to watch the entire movie by a human expert is the most time consuming step. This paper discusses the problem of missing transcription, where the subtitle blocks corresponding to some speech segments in the DEC are non-existent. We present a solution to augment human correction process by automatically identifying the timings associated with the non-transcribed dialogues in a language agnostic manner. The correction step can then be performed by either human-in-the-loop mechanism or automatically using neural transcription (speech-to-text in same language) and translation (text-to-text in different languages) engines. Our method uses a language agnostic neural Voice Activity Detector (VAD) and an Audio Classifier (AC) trained explicitly on DEC corpora for better generalization. The method consists of three steps: First, we use VAD to identify the timings associated with dialogues (predicted speech blocks). Second, we refine those timings using the AC module by removing the timings associated with the leading and trailing non-speech segments identified as speech by VAD. Finally, we compare the predicted dialogue timings to the dialogue timings present in the subtitle file (subtitle speech blocks) and flag the missing transcriptions. We empirically demonstrate that the proposed method a) reduces incorrect predicted missing subtitle timings by 10%, b) improves the predicted missing subtitle timings by 2.5%, c) reduces False Positive Rate (FPR) of overextending the predicted timings by 77%, and d) improves the predicted speech block-level precision by a 119% over VAD baseline on a human-annotated dataset of missing subtitle speech blocks.

Keywords: Digital Entertainment Content (DEC); Missing Subtitles; Voice Activity Detection (VAD); Audio Classifier (AC); Deep Learning; Audio Processing

1 Introduction

Content localization is fundamental to DEC expansion into newer territories and enhancement of viewing experience. Subtitling or creation of subtitles is a vital component of content localization. Subtitles are composed of the dialogues and their associated timings; known as subtitle speech blocks and plot pertinent non-speech sounds along with their timings; known as captions. We infer the timings associated without dialogues or with captions as subtitle non-speech blocks. Subtitling is a manual process which includes linear watch of a title, identification of timestamps associated with dialogues and transcription of the dialogues followed by translation into the target language. This process results in errors such as missing transcriptions (missing subtitle speech blocks), out-of-sync subtitle blocks with the audio and incorrect translations. These erroneous subtitles result in an unpleasant viewing experience and negatively affect the viewership. This paper focuses on the missing subtitle blocks error that significantly affects the subtitle quality. Based on data collected as per our internal Language Quality Program (LQP), for a random subset of 100 subtitles submitted to our system by third party linguistic experts, $\approx$ 1% of them contain one or more missing subtitle speech blocks, making it one of the largest problems related to subtitle localisation. Missing subtitle blocks

*Correspondence: ghoney@amazon.com
Prime Video International Expansion, Amazon, India, Bangalore, India
Full list of author information is available at the end of the article

occur due to, a) non-transcribed foreign language spoken in a dialogue, b) human errors in creating the subtitles and c) inadequate quality checks post the subtitle creation.

Catalog expansion and multi-lingual nature of audio and subtitle pairs require an automated and language agnostic approach to detect missing subtitle blocks. Identifying missing subtitle blocks is a manual process, which requires a linear watch of the title by a linguistic expert who identifies the timestamps and fills the missing text. Identification of timestamps contributes for the greatest time ($\approx$ 90%) in the process. Also, there exists multiple subtitles and audio tracks across several languages for a given title. Therefore, missing subtitle block detection is a costly and time consuming process. Hence, we propose an automated solution to identify the timestamps associated with the missing subtitle speech blocks using a language agnostic Voice Activity Detection (VAD) and Audio Classification (AC) model. The language agnostic characteristic of VAD removes the dependency on a linguistic expert and significantly reduces the time taken for missing block detection in the titles by reducing manual touch points. Once the missing timings are identified, we can either use an Automated Speech Recognition (ASR) engine or a linguistic expert/creative director to transcribe and translate the audio corresponding only to the missing timestamps.

For a given DEC title, we detect missing subtitle blocks by identifying the time stamps associated with speech segments in the audio and matching them with the time stamps present in the respective subtitle file. A given DEC title can be localized across multiple languages and can contain some dialogues spoken in a language which is different from its native locale. Hence, we use a language agnostic VAD to identify timings associated with dialogues. However, a typical VAD model can lead to various False Positive (FP) cases such as a) contextual background noises like traffic noises, crowd noises, music etc., and b) atypical speech patterns like whispering, shouting, singing, and electronic voices. To reduce the number of falsely identified missing blocks, we fine tune the VAD's predicted timings using an Audio Classification model. We evaluate the performance of the missing subtitle block detector on a synthetic and a human annotated corpora consisting of missing subtitle speech blocks.

The main contributions of this paper are as follows: First, we propose a language agnostic approach for missing subtitle block detection using VAD and AC models. Our approach alleviates the dependencies on language reliant systems such as Automatic Speech Recognition (ASR) and text translation models for this task. Second, we use a VAD model explicitly trained on DEC corpus, enhancing the robustness of the proposed method to various background noises present in DEC titles. Third, we present a baseline solution using the neural VAD model. Fourth, despite its robustness, the VAD system potentially identifies certain sounds as human speech. The effect of such false positives is reduced by our multiclass AC model, which identifies 121 categories of sounds and is trained on DEC and open source corpora. Finally, we show that our model results in; a) 10% reduction in incorrect predicted missing subtitle timings, b) 2.5% improvement in identifying the correct locations of missing subtitles on real-world dataset, c) 77% reduction in False Positive Rate (FPR) of overextending the predicted speech timings, and d) 119% improvement in the predicted speech block-level precision over a VAD baseline on a real-world human-annotated dataset of missing subtitle speech blocks.

2 Related works

In this section, we briefly discuss the literature related to voice activity detection and audio classification as they form the key components of our proposed method.

Voice Activity Detection: Recently, there has been tremendous progress in deep learning for sequences, especially for VAD in DEC. Mateju *et al.* [1] used a deep neural network trained on noise augmented dataset along with smoothing of the output for speech activity detection in movies. Jang *et al.* [2] used a 2 layered DNN with MFCC as the input feature for VAD in movies. Zhang *et al.* [3] used boosted deep neural network bDNN that generated multiple predictions from different contexts of a single frame by only one DNN and then aggregated the predictions for a better prediction of the frame. Hwang [4] used ensemble of DNNs. Kang *et al.* [4] used Multi Task Learning (MTL) with DNN to estimate clean features from noisy features as well as VAD probabilities.

Audio category classification: Audio classification predicts the audio tags in an audio clip. Convolutional Neural Networks (CNNs) have been used [5] to predict the tags of audio recordings. CNN based systems have achieved state-of-the-art performance in several DCASE challenge tasks including acoustic scene classification [6] and sound event detection [7]. A milestone for audio pattern recognition was the release of AudioSet [8], a dataset containing over 5,000 hours of audio recordings with 527 sound classes. Several CNN based models have been proposed for large scale audio classification [9–13], however, Pretrained Audio Neural Network or PANN [14] is a VGGish [15] CNN based model that achieves the state-of-the-art

Table 1 DEC-1100 video distribution by language, where the language code is identified using ISO-639 [19] (639-1) nomenclature.

Language Code	en	de	hi	ja	ko	fr	te	ta	es
Percentage	68	1	13	13	2	1	1	1	1

result for Audioset classification task. In the next section, we present the approach to detect the missing subtitle speech blocks.

3 Methodology

The proposed approach to identify missing subtitle speech blocks involves two steps: 1) identification of speech and non-speech duration using VAD and 2) improvement of these duration through removal of false positive cases using AC model. In this section, First, we describe the VAD and AC model architectures. Second, the datasets used to train and validate them. Third, comparison with their corresponding state-of-the-art models which justifies our architectural design choices and Fourth, the method for missing subtitle detection using these models.

3.1 Voice Activity Detection Model (VAD)

A VAD model trained on domain specific DEC dataset consisting of several languages and background noises results in better generalization and language agnostic characteristic compared to the models trained on non-DEC focused datasets [16]. Therefore, we use an in-house developed Gated Recurrent Unit (GRU) [17,18] based VAD trained on a proprietary DEC dataset (DEC-1100). This dataset consists of 1100 proprietary videos (≈ 450 hours) along with their subtitles spanning 9 languages and 5+ genres (Action, Comedy, Documentary, Drama, Animation, etc.) making it one of the largest DEC based dataset used to train the VAD model. Table 1 presents the language distribution of the dataset.

Train, Validation and Test set creation: To create the training set, we divide the videos into 800 milliseconds (ms) non-overlapping clips and label them into speech and non-speech using the timing information in the subtitles. This results in 1.1 M speech and non-speech clips respectively. Similarly, the validation set consists of 0.1 M speech and non-speech clips respectively. The test set consists of human validated 18k and 27k speech and non-speech clips respectively. It is curated from 33 movies which are not part of the training and validation sets (DEC-1100).

We use the value of 800 ms for two reasons. First, a human speech block in a timed-text file should persist on the screen for a minimum duration between $5/6^{\text{th}}$ of a second to one second, as recommended by several industry standard guidelines [20]. These guidelines are based on the studies conducted on the reading speed of viewers. Second, disambiguation of a clip below 500 ms into speech and non-speech is difficult for human evaluators based on our manual inspection of clips.

Table 2 Measures compared with various VAD models trained on DEC-1100 dataset and tested on DEC based human annotated test set

Model Name	Accuracy	AUC	Precision	Recall	F-score
CLDNN	0.852	0.915	0.877	0.852	0.854
CNNTD	0.866	0.947	0.876	0.866	0.867
GRU	0.871	**0.951**	**0.887**	0.871	0.872
STNET	0.863	0.940	0.863	0.863	0.861
TCN	**0.875**	0.900	**0.887**	0.875	0.876
WebRTC	0.615	-	0.757	0.615	0.597

VAD Model: The network diagram of the VAD model is shown in the Figure 1. The model is a modification of the LSTM based VAD model described in [18]. It consists of two parallel bidirectional GRUs each containing two layers of 128 dimension each. The outputs of the GRUs are time weighted, concatenated and passed through two Fully Connected (FC) layers of 128 and 2 dimensions respectively, followed by a softmax. The model takes a one dimensional audio sequence of length 800 ms sampled at 32 kHz as input, generates time-frequency based features and returns the probability of speech. The feature extraction module converts the audio clip in two feature maps namely, the magnitude Short Term Fourier Transform (STFT) with 54 time bins and 128 frequency bins and the frequency based 128 dimensional reassigned frequency or Instantaneous Frequencies (IF) [21] with 54 time bins. IFs were proposed as a feature by Longbiao *et al.* and Iain *et al.* [22,23] and have shown to improve VADs performance. The magnitude STFT and IFs are calculated using a 25 ms window (800 samples) and 10 ms (320 samples) hop length.

Results: The GRU based VAD either outperforms or performs at-par with several state-of-the-art neural models such as Temporal Convolution Network (TCN) [24], Convolutional and Self Attention (STNET) [25] transformer encoder based network [26], VGG-net based Time Distributed CNN (CNNTD) [16], raw audio waveform based CLDNN [27] and webRTC VAD [28] in terms of Area Under Curve (AUC), precision, recall and F-scores. Table 2 presents the results for various VAD models trained on DEC-1100 dataset and tested on DEC based human-annotated test set. We now describe our Audio Classification Model.

Neural VAD model has a false positive rate of ≈ 15% and tags sounds such as songs and unintelligible human sounds like sighs, grunts, laughs, cry as human speech. Therefore, we use the AC model to remove these false positives, which is described in the following subsection.

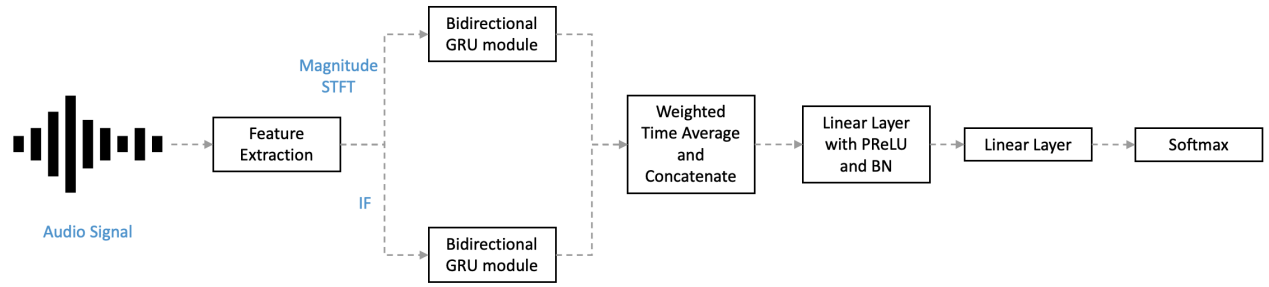

Figure 1 Network architecture of our GRU based VAD model. The model uses 800 ms audio signal sampled at 32 kHz as input, extracts magnitude (STFT) and Instantaneous Frequency (IF) spectrograms using feature extraction module. These two spectrograms are then normalized using Batch Normalization (BN) and passed through two parallel two layered Bi-GRU module. The outputs of the GRUs are time averaged, concatenated and passed through linear layer (128 dimension) followed by a Parametric ReLU (PReLU), Batch Normalization (BN), linear layer (2 dimensional) and a softmax to generate probability of speech and non-speech.

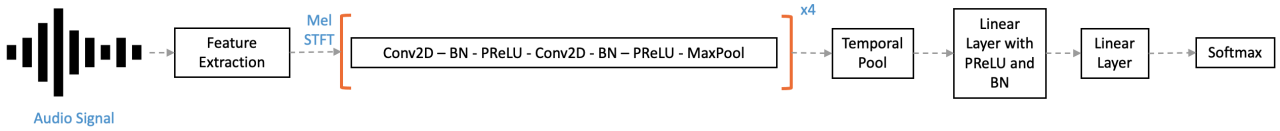

Figure 2 Network architecture of our Audio Classification model. The network takes in 2 s audio clips sampled at 48 kHz as input and extracts log mel spectrogram as input. The spectrogram is passed as an input to the VGGish network consisting of 4 covolutional blocks. Each block consists of conv2D-BatchNorm(BN)-PReLU-conv2D-BN-PReLU and a 2×2 MaxPool2D layer. Following the blocks, we pool along the temporal axis and reshape the input into a 2D array. This input is passed through two fully connected layers of sizes 512 (with a dropout of 0.5) and 121 respectively. Finally, we perform a softmax on 121 categories.

3.2 Audio Classification Model (AC)

We trained a generic Audio Classifier (AC) to detect presence of captions and the audio events falsely classified as speech by the VAD model. This model is trained on an audio event dataset consisting of 121 different human annotated sound event clips from DEC-1100 dataset, 1800 videos from another internal proprietary repository (known as DEC-1800) and two publicly available datasets namely, FSDKaggle2019 [29] and Google Audioset [8]. We now describe the training and testing dataset creation process.

Train, Validation and Test set creation: We use the time duration of captions from DEC-1100 and DEC-1800 to create the multi-class dataset. We categorize these sounds into 121 categories. The categories includes human sounds (grunts, sigh, laugh, cough etc.), music and instrument related sounds (chant, song, background music, jingle etc.), animal sounds (bark, meow etc.), machine sounds (traffic, gunshots etc.) and other environmental sounds (wind, waves etc.). The categories are outlined below:

`applause, bang, bark, beep, blare, bleat, breathe heavily, burp, buzz, chant, chatter, cheer, chime, chirp, clank, clap, clatter, clear throat, click, clink, cluck, coo, cough, crack, crackle, crash, creak, croak, cry, dial, ding, door_or_drawer_open_or_close, drill, echo, engine, exclaim, exhale, explosion, fart, flapping, footstep, gasp, groan, growl, grumble, grunt, gunfire, helicopter, hiccup, hiss, honk, howl, hum, inhale_or_exhale, instrument-play, jingle, knock, laugh, meow, moan, moo, mosquito, muffle, music, mutter, neigh, noise, not_a_caption, oink, others, pant, pop, quack, rain,rattle, revving, ring, roar, rumble, rustle, scoff, scream, screech, shatter, shiver, sigh, silence, siren, sizzle, snap, snarl, sneeze, sniff, snore, snort, sob, song, spit, squawk, squeak, squeal, static, talk, thud, tick, toll, tone, traffic, trill, type, water run, waves, whimper, whine, whirr, whisper, whistle, whoop_or_whoosh, wind, yell, yelp, others, not_a_caption, silence.`

Finally, we extract the audio segments from their corresponding caption timings present in the subtitle file. Similarly, we extract the segments from the two public datasets with the above mentioned categories. We divide the segments from both public and

Table 3 Performance comparison of various audio classification methods on human labelled test set.

Model	Accuracy	AUC	Precision	Recall	F-score	Average Recall	Top3 Accuracy
GRU	54.7%	0.972	53.8%	54.7%	53.7%	63.7%	76.1%
ResNeXt	63.8%	0.984	63.4%	63.8%	63.3%	73.1%	83.4%
CNNTD-small	67.4%	0.9867	66.7%	67.4%	66.8%	74.8%	85.9%
CNNTD-large	71.8%	**0.9876**	70.9%	71.8%	71.0%	**77.1%**	**88.5%**
PANNs	**73.22%**	0.9546	**72.75%**	**73.22%**	**72.88%**	58.31%	88.41%

Figure 3 CNNTD-small and CNNTD-large model architectures.

proprietary datasets into 2 s clips with 50% overlap between consecutive clips. We choose a duration of 2 s due to two reasons: First, 90% of the captions duration in DEC-1100 and DEC-1800 are smaller than 2.3 seconds. Second, several sounds such as 'instrument-play', 'songs', 'chant','echo' etc., require longer time duration for classification as compared to VAD. The distribution of audio clip-label pair in the resulting dataset is as follows: a) DEC-1100: 51,337, b) DEC-1800: 90,333, c) FSDKaggle2019: 1,51,989 and d) Google Audioset: 9,354.

Further, we perform human annotation where each clip was tagged by 2 annotators to minimize the human error. We retain the clips which had agreement between the two annotators resulting in 2,00,000 clips sampled at 48 kHz. We extract log scaled mel-STFT of the clips with 128 bins and 134 time frames using a window size of 25 ms (1200 samples) and hop length of 15 ms (720 samples). We use 80 % of this dataset for training, 10% for validation and remaining for testing purpose. However, we observe a data imbalance of 7500x between the samples of largest and smallest category. Hence, we use an approach similar to Spec-Augment [30] for synthesizing the training samples of the imbalanced classes using the following four techniques: First, time warping of spectrogram by a factor between 0.8 and 1.2 of the spectrogram's time bins. Second, time and frequency stretching by a random factor between 0.8 and 1.2 of the spectrogram's time and frequency bins. Third, global spectrogram magnitude shift in both positive and negative directions by a random factor between 0.05 and 0.1 of the mean amplitude. and Fourth, introducing time-frequency masking by random masking 20% continuous time and frequency bins. This process results in 1.5M training samples across 121 classes.

AC Model: The network diagram of the AC model is shown in the Figure 2. The AC is a VGGish model known as CNNTD [16] that consists of 4 convolutional blocks of 2 layers each followed by temporal pooling (TP) and two FC layers followed by a softmax over 121 categories. We explore two variants of the VGGish model; a) CNNTD-large: with 13 M parameters and b) CNNTD-small with 2.9M parameters as shown in the figure 3.

Results: We compare the models against PANNs [14], ResNeXt [31] and GRU based [32, 33] models. Comparison results for these methods can be found in the table 3. We observe that CNNTD-large model results in the best AUC, average recall and top3 accuracy among all the models. Hence, we use CNNTD-large model as our AC model to be used as a component of missing subtitle detector. In the following subsection, we describe the approach to detect missing subtitle speech blocks using the GRU based VAD and VGGish CNNTD-large AC model.

3.3 Missing Subtitle block detection using VAD and AC models

Our proposed method consists of 3 stages, as depicted in the Figure 4. First, we obtain the timings of speech/non-speech segments or blocks from VAD and AC models independently. Second, we merge the two timings and remove the false positives of VAD. Finally, we compare the predicted timings with the timings in the subtitle file and identify the positions of missing speech in the file. We now describe the timing generation process using the two models.

VAD Inference: The VAD inference consists of six steps: First, we extract audio from the video and divide it into 800 ms clips with 90% overlap between consecutive clips. Second, we use VAD model to obtain the probability of speech for each 800 ms clip. Third, due to overlap of 90% between clips, we assign the probability of first 800 ms clip to first 80 ms segment, assign

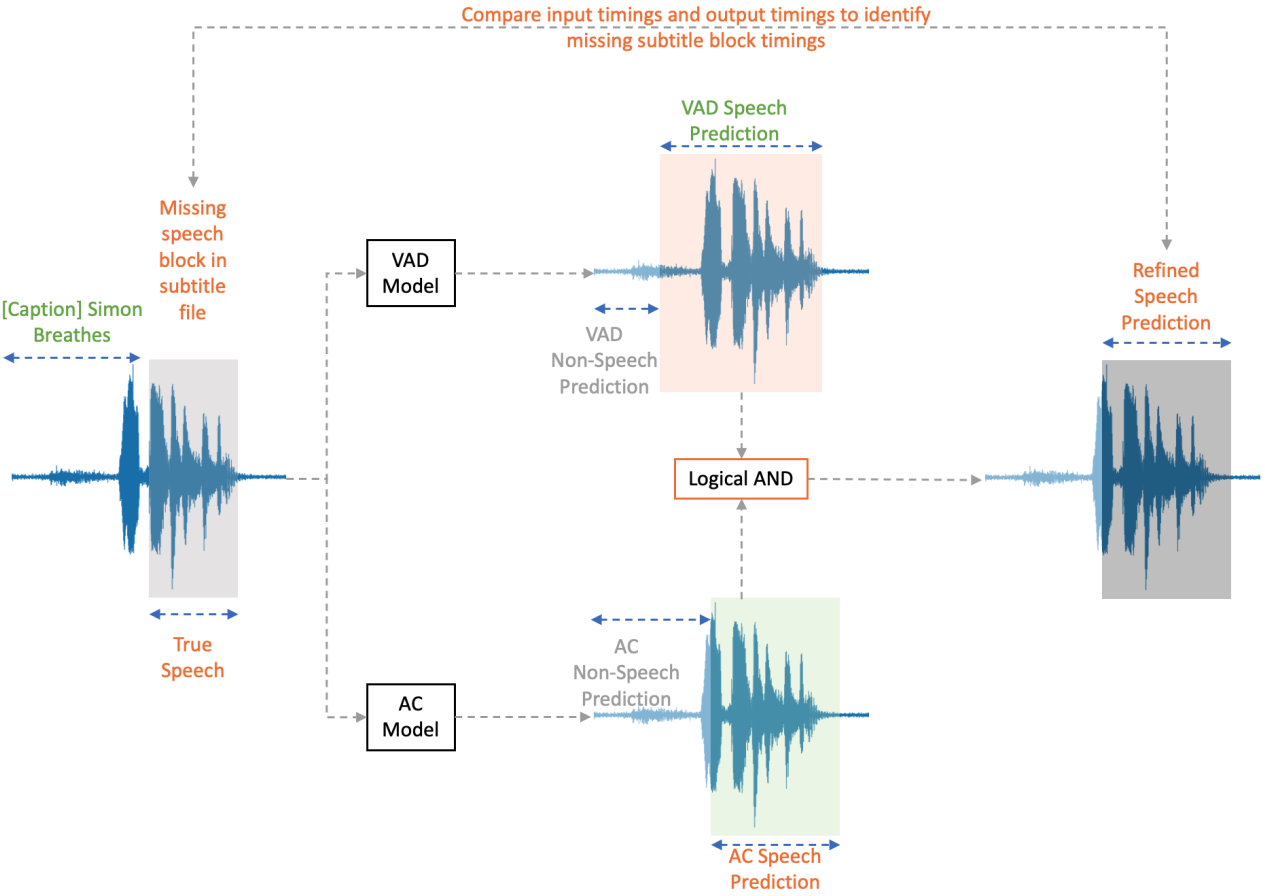

Figure 4 Proposed method for combined VAD and AC inference and the algorithm to identify missing subtitle blocks. Consider the dialogue at the start which consists of a caption (Simon Breathes) and speech following it. However, this dialogue is missing from the subtitle file. To identify the true speech timings, we divide the audio in 800 ms (with 90% overlap) and 2 s clips (with no overlap) and pass them to VAD and AC models respectively. Following the VAD and AC timing generation step for the clips, we perform a logical AND between the timings and generate the refined predicted speech blocks. VAD can potentially identify the caption (Simon Breathes) as a speech block. The time duration associated with the caption is identified by the AC model and is removed from the VAD's timing to generate the correct timings. We then compare the timings of predicted refined speech block to the timings present in the subtitle blocks and predict the missing subtitle blocks.

the probability of second 800 ms clip to second 80 ms segment and so on. Fourth, to filter spurious probabilities, we smooth the resulting probability vector using a moving average window of length 35. Fifth, we join the consecutive 80 ms segments having probability > 0.5 to form speech blocks and obtain their timings. Finally, we combine the consecutive speech blocks where end of the former and start of the latter segment is less than 300 ms to obtain final VAD speech blocks. We merged the blocks that are < 300 ms apart because significant pauses associated with commas, blanks, punctuations are around 300 ms. We chose the window length and the probability threshold through a hyperparameter tuning step. The VAD inference steps are depicted in the Figure 5.

AC Inference: The AC inference consists of 4 steps: First, we divide the extracted audio into 2 s clips without overlap. Second, we obtain the probability of various categories from the AC model for a given clip. Third, we identify top-K (K = 3) categories and consider the clip as non-speech if it contains any of the following with a probability $p \geq 0.6$: 'music', 'song', 'instrument-play', 'groan', 'inhale_or_exhale', 'sigh', 'clear throat', 'breathe heavily', 'grunt', 'cough', 'gasp' and 'exhale'. These categories were chosen on the basis of most frequent captions present in DEC-1100 and DEC-1800. We make a simplifying assumption about other categories and consider the rest as speech. Finally, we combine the consecutive speech

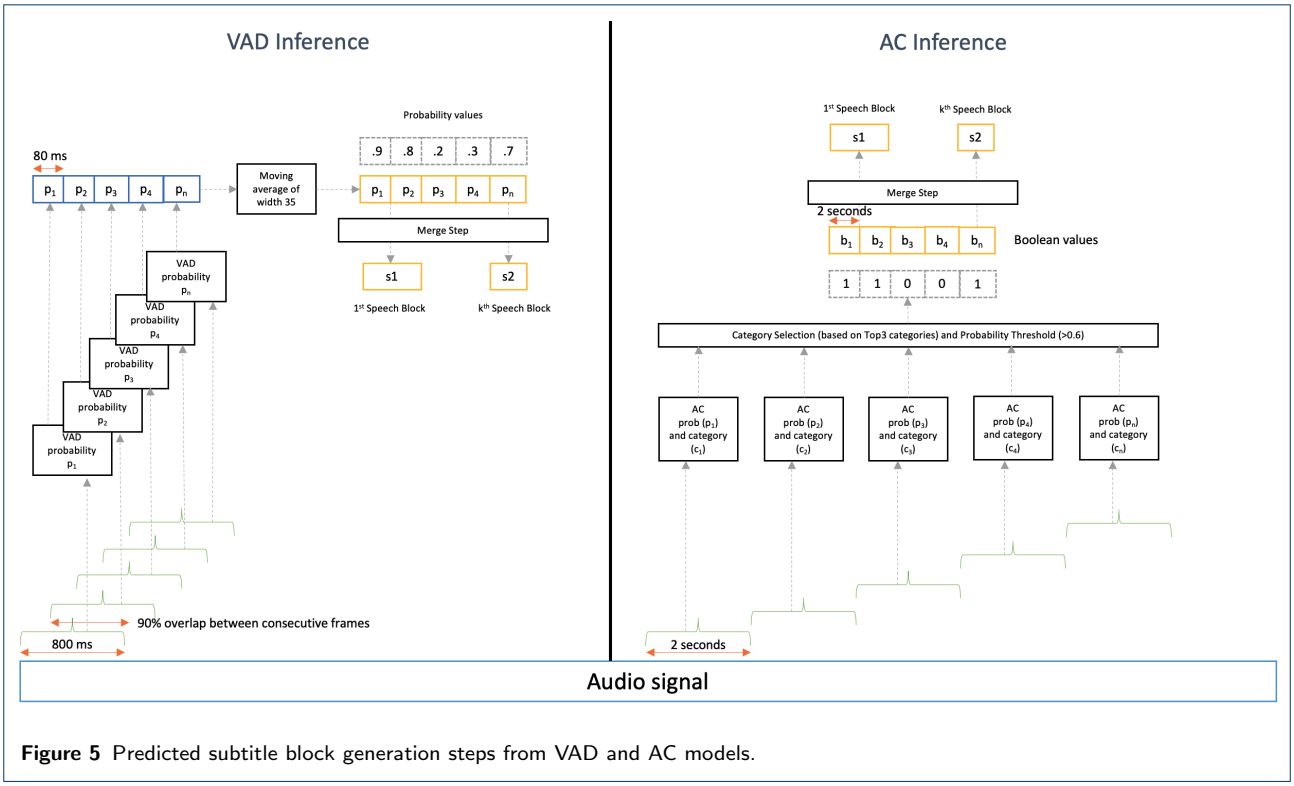

Figure 5 Predicted subtitle block generation steps from VAD and AC models.

segments to form AC speech blocks and obtain their timings. AC inference steps are depicted in the Fig. 5.

Combining the predictions: We create two binary arrays of length equal to the length of the audio in milliseconds (ms) using the predictions of the above two steps respectively. Since the VAD and AC models work at different granularity, we use 1 ms as a scale for the final array to enable easier extrapolation and merging. We extrapolate the predictions of VAD and AC models to ms level and fill the arrays with '1' at speech locations and rest with '0'. Subsequently, we perform a logical AND operation between the two arrays. Finally, we obtain the timings of speech and non-speech blocks by combining the consecutive predictions.

Identification of missing blocks: We compute the overlap between every predicted speech block's timings with the speech block's timings in the subtitle file. We consider a predicted speech block 'covered', if it overlaps with a subtitle speech block for more than $t = 800$ ms. During inference, if a predicted speech block is not covered by any subtitle speech block, we consider the block as missing from the subtitle file. In the following section, we outline the datasets, metrics and comparison results on two DEC based missing subtitle block datasets.

4 Experiments And Results

In this section, we present the datasets, metrics, hyperparameter tuning and results of our experiments. We use VAD model as the baseline to benchmark the proposed method. Further, we also compare against the speech timings obtained from the proprietary language dependent neural ASR model similar to model used by Kaldi ASR [34, 35].

4.1 Datasets

Owing to lack of publicly available datasets on the problem, we use two proprietary datasets in our evaluations. First, we create a synthetic dataset of missing subtitles from 50 proprietary videos sampled from Amazon Originals. These videos consists of synced subtitles in English language. To create the dataset, we randomly remove 10% of the subtitle speech blocks and treat them as missing subtitle blocks. Second, we use dataset of 430 incorrectly synced DEC video-subtitle pairs that contains missing subtitle blocks obtained through our internal Language Quality Program. We used human validation to identify 354 missing speech blocks with time duration > 500 ms.

4.2 Metrics

We use two metrics to evaluate the performance of the models: 1) Subtitle block duration based metric - *coverage* and 2) Subtitle block detection based metrics - *False Positive Rate (FPR)*, *Precision* and *Recall*.

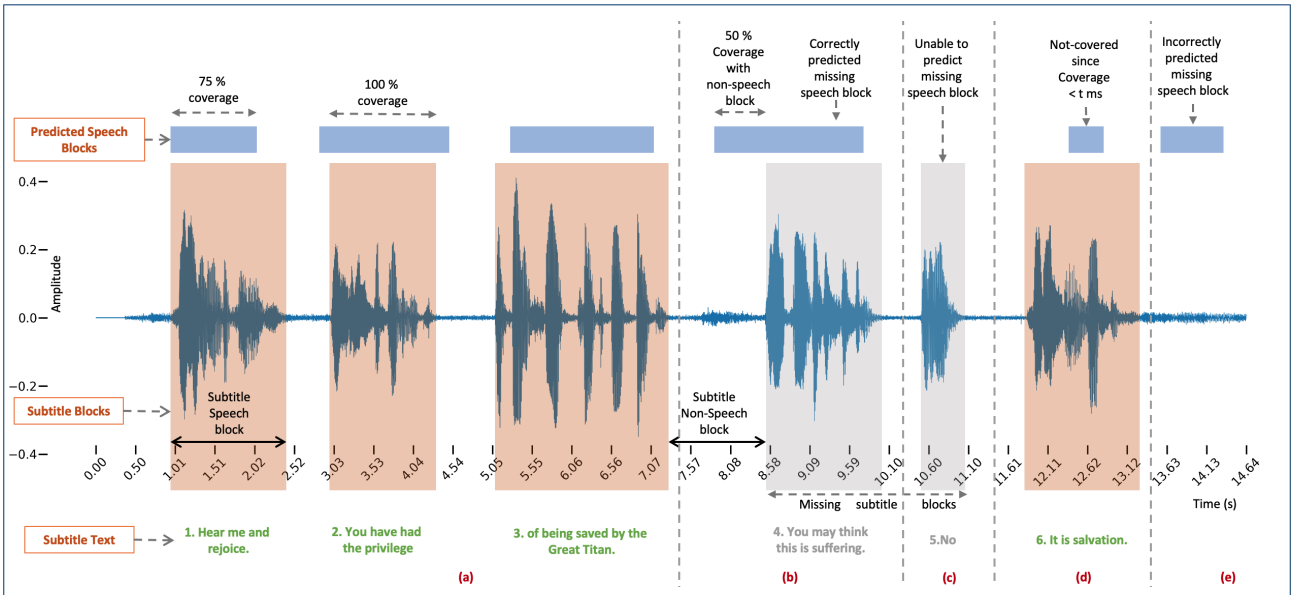

Figure 6 The figure depicting subtitle speech blocks in peach (overlaid on audio track) in the middle, predicted speech blocks in blue at the top and subtitle text at the bottom. The figure highlights several output cases of our algorithm: (a) depicts the coverage of predicted speech blocks (blue) with the subtitle speech blocks (peach), (b) shows the subtitle non-speech block, missing subtitle speech block (in light grey) and predicted speech block that correctly predicts the missing subtitle location but overlaps with the non-speech segment as well, (c) our algorithm is unable to predict the missing speech block, (d) our algorithm makes a prediction with coverage $< t$ ms and hence, fails to detect the speech block, (e) the algorithm falsely identifies a non-speech region as speech.

While duration based metric provide the effectiveness of identifying the correct timings of missing blocks, the block level metrics identifies the effectiveness in identifying the missing blocks themselves. Coverage [36] is defined as the ratio of intersection duration of the hypothesis segment with reference segment and the duration of reference segment.

Coverage: We calculate the coverage metric across two terms: First, between the predicted speech blocks (hypothesis) with the missing speech blocks in the subtitle file (reference). We term predicted speech blocks with intersection $t > 800$ ms with the reference missing speech blocks as *correctly predicted missing speech blocks* (Figure 6, part (a,b)). On the other hand, *incorrectly predicted speech blocks* have a intersection $t > 800$ ms with non-speech blocks and are without intersection with the missing speech blocks in the subtitle file (Figure 6, part (e)). Second, for every correctly predicted missing speech block (hypothesis) we compute its intersection with neighbouring non-speech blocks in the subtitle file (reference). The first value indicates the effectiveness of method to correctly predict the time duration of the missing subtitle blocks. The second value highlights the bleeding of predicted missing speech time duration into non-speech regions.

FPR, **Precision and Recall**: These metrics quantify the efficacy of the method in detecting missing speech blocks. First, we compute the FPR that quantifies the percentage of correctly predicted missing speech blocks that over-extends to non-speech blocks of the subtitle file. The FPR is computed in two steps: First, we identify the number of correctly predicted missing speech blocks that also intersects with the neighbouring non-speech subtitle blocks and Second, we take their ratio with the total number of non-speech subtitle blocks. Next, we compute the precision as the ratio of the number of correctly predicted missing speech blocks to the total number of predicted speech blocks. Finally, we compute the recall as the ratio of the number of correctly predicted speech blocks to the total number of missing subtitle blocks.

4.3 Comparison

In this section, we present the duration based and block-level based analysis on our synthetic and real world missing subtitle datasets.

Analysis on synthetic dataset: Table 4 presents the coverage percentages of using: (a) VAD baseline, (b) VAD + AC and (c) proprietary ASR for determining the missing subtitle blocks. For the VAD baseline and proprietary ASR, a procedure similar to Section 3.3 was followed to flag the missing subtitle blocks. This included forming speech segments using the predicted probabilities and calculating overlap with the subtitle speech blocks to flag the missing segments. We observe that the VAD baseline model results in $\approx 82\%$

Table 4 Analysis on Synthetic dataset

Segment Level Metric	VAD	VAD + AC	ASR
Coverage between predicted speech to reference missing speech sections	**81.8**	79.37	74.21
Coverage between predicted speech to reference non-speech sections	15.04	**12.42**	26.93

Table 5 Analysis on Human Annotated dataset

Segment Level Metric	VAD	VAD + AC	ASR
Coverage between predicted speech to reference missing speech sections	72.15	**74.91**	60.53
Coverage between predicted speech to reference non-speech sections	45.22	**35.7**	39.3

Table 6 Block-level analysis on Human Annotated dataset. ↓: Lower is better and ↑: higher is better.

Threshold (t in ms)	FPR ↓			Precision ↑			Recall ↑		
	VAD	VAD +AC	% Reduction	VAD	VAD + AC	% Improvement	VAD	VAD + AC	% Improvement
300	0.214	**0.106**	50.467	0.277	**0.438**	58.123	**0.712**	0.69	-3.090
500	0.195	**0.072**	63.077	0.298	**0.552**	85.235	**0.707**	0.667	-5.658
800	0.169	**0.039**	76.923	0.334	**0.732**	119.162	**0.702**	0.631	-10.114
1000	0.149	**0.03**	79.866	0.369	**0.819**	121.951	**0.695**	0.597	-14.101
1200	0.134	**0.024**	82.090	0.398	**0.853**	114.322	**0.687**	0.558	-18.777

coverage with reference missing subtitle blocks. However, predicted speech coverage with reference non-speech blocks from the subtitle file is close to 15%. This happens as VAD falsely identifies some non intelligible human sounds and music categories as human voice. Using the AC model, we are able to bring the predicted speech coverage with reference non-speech blocks down by 2.5% from the VAD baseline, but at the cost of 2% reduction in coverage with reference missing subtitle blocks. The ASR system which was not trained on DEC dataset results in very high predicted speech coverage ($\approx$ 27%) with reference non-speech blocks.

Analysis on human annotated dataset: From table 5, we observe that VAD + AC model outperforms VAD baseline and ASR in terms of coverage. The ASR system has low coverage with the reference missing subtitle blocks mainly due the presence of noise in the video clips. The VAD + AC model model significantly reduces the percentage of predicted speech coverage with reference non-speech blocks ($\approx$ 10%) as compared to the VAD baseline approach and improves upon the predicted speech coverage with reference missing speech blocks (by $\approx$2.5%). The large value of predicted speech coverage with reference non-speech blocks is mainly due to: a) incorrect timing annotation, and b) songs being identified as speech by all three models, as verified through a manual inspection of the falsely predicted speech segments.

Table 6 presents the block-level performance of the baseline VAD model and our proposed VAD + AC method on the human annotated dataset as detection threshold t is varied while predicting the missing-subtitle blocks. Here, we do not compare ASR performance as VAD and VAD + AC models are empirically observed to perform better than ASR system. We observe that our proposed VAD + AC model outperforms the VAD baseline by a significant margin in terms of FPR and Precision. Results indicate that as the detection threshold increases, the FPR value of both the VAD and VAD + AC models reduces significantly as the models become more confident in predicting the missing subtitle blocks. The FPR value of VAD + AC model is much lower than VAD baseline as AC model reduces the effect of incorrect predictions of VAD. At $t = 800$ ms which is the input duration for VAD, the VAD + AC results in $\approx$ 77% reduction in FPR.

Similarly, VAD + AC significantly outperforms the VAD baseline in terms of precision. At $t = 800$ ms, the VAD + AC model results in 119% increase in precision as compared to its VAD counterpart by removing the false detections. However, the VAD + AC model results in a 10% reduction in recall at $t = 800$ ms which is marginal reduction as compared to VAD baseline. This reduction occurs as AC model has the potential to remove certain true speech segments present in VAD due to its input length threshold of 2 seconds.

5 Conclusions

We proposed two automated language-agnostic methods for missing subtitle detection. We showed that a VAD can be suitably used for detecting audio segments having a missing subtitle blocks. Further, conjugating the VAD model with an AC model improves the detection by effectively reducing the false positive cases of VAD. We presented a performance comparison on two DEC missing subtitle blocks datasets and showed that our proposed method works significantly well for the task at hand. Our proposed method is language agnostic and achieves an true coverage of 75% on a human-annotated dataset and a configurable block-level precision of upto 0.85. The proposed approach can also be reasonably applied to other VAD methods proposed for various applications apart from missing

subtitle detection. Since our method reduces the false-positives of the VAD model, it can be extended to other use-cases such as speech identification or subtitle drift detection to reduce the false-positive cases of the VAD model.

Acknowledgements
Not applicable

Funding
Not applicable

Abbreviations
Not applicable

Availability of data and materials
The data used in the studies presented in this paper is proprietary and cannot be released publicly.

Ethics approval and consent to participate
Not applicable

Competing interests
The authors declare that they have no competing interests.

Consent for publication
The authors provide their consent to publish the manuscript after acceptance.

Authors' contributions
We will update this section during the camera-ready submission.

Authors' information
Not applicable

Author details
Prime Video International Expansion, Amazon, India, Bangalore, India.

References

1. Mateju, L., Cerva, P., Zdánský, J., Málek, J. Speech activity detection in online broadcast transcription using deep neural networks and weighted finite state transducers. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017, pp. 5460–5464 (2017). doi:10.1109/ICASSP.2017.7953200. https://doi.org/10.1109/ICASSP.2017.7953200
2. Jang, I., Ahn, C., Seo, J., Jang, Y. Enhanced feature extraction for speech detection in media audio. In: Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017, pp. 479–483 (2017). http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0792.html
3. Zhang, X., Wang, D. Boosting contextual information for deep neural network based voice activity detection. IEEE ACM Trans. Audio Speech Lang. Process. **24**(2), 252–264 (2016). doi:10.1109/TASLP.2015.2505415
4. Hwang, I., Park, H., Chang, J. Ensemble of deep neural networks using acoustic environment classification for statistical model-based voice activity detection. Comput. Speech Lang. **38**, 1–12 (2016). doi:10.1016/j.csl.2015.11.003
5. Choi, K., Fazekas, G., Sandler, M.B. Automatic tagging using deep convolutional neural networks. In: Mandel, M.I., Devaney, J., Turnbull, D., Tzanetakis, G. (eds.) Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7-11, 2016, pp. 805–811 (2016). https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/009_Paper.pdf
6. Mesaros, A., Heittola, T., Virtanen, T. A multi-device dataset for urban acoustic scene classification. CoRR **abs/1807.09840** (2018). 1807.09840
7. Cakir, E., Heittola, T., Huttunen, H., Virtanen, T. Polyphonic sound event detection using multi label deep neural networks. In: 2015 International Joint Conference on Neural Networks, IJCNN 2015, Killarney, Ireland, July 12-17, 2015, pp. 1–7 (2015). doi:10.1109/IJCNN.2015.7280624. https://doi.org/10.1109/IJCNN.2015.7280624
8. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017, pp. 776–780 (2017). doi:10.1109/ICASSP.2017.7952261. https://doi.org/10.1109/ICASSP.2017.7952261
9. Kong, Q., Xu, Y., Wang, W., Plumbley, M.D. Audio set classification with attention model: A probabilistic perspective. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018, pp. 316–320 (2018). doi:10.1109/ICASSP.2018.8461392. https://doi.org/10.1109/ICASSP.2018.8461392
10. Kong, Q., Yu, C., Xu, Y., Iqbal, T., Wang, W., Plumbley, M.D. Weakly labelled audioset tagging with attention neural networks. IEEE ACM Trans. Audio Speech Lang. Process. **27**(11), 1791–1802 (2019). doi:10.1109/TASLP.2019.2930913
11. Darna-Sequeiros, J., Toledano, D.T. Audio event detection on google's audio set database: Preliminary results using different types of dnns. In: Luque, J., Bonafonte, A., Pujol, F.A., Teixeira, A.J.S. (eds.) Fourth International Conference, IberSPEECH 2018, Barcelona, Spain, 21-23 November 2018, Proceedings, pp. 64–67 (2018). doi:10.21437/IberSPEECH.2018-14. https://doi.org/10.21437/IberSPEECH.2018-14
12. Verbitskiy, S., Vyshegorodtsev, V. Eranns: Efficient residual audio neural networks for audio pattern recognition. CoRR **abs/2106.01621** (2021). 2106.01621
13. Hershey, S., Ellis, D.P.W., Fonseca, E., Jansen, A., Liu, C., Moore, R.C., Plakal, M. The benefit of temporally-strong labels in audio event classification. CoRR **abs/2105.07031** (2021). 2105.07031
14. Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE ACM Trans. Audio Speech Lang. Process. **28**, 2880–2894 (2020). doi:10.1109/TASLP.2020.3030497
15. Simonyan, K., Zisserman, A. Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015). http://arxiv.org/abs/1409.1556
16. Hebbar, R., Somandepalli, K., Narayanan, S. Robust speech activity detection in movie audio: Data resources and experimental evaluation. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4105–4109 (2019). IEEE
17. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
18. Eyben, F., Weninger, F., Squartini, S., Schuller, B. Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 483–487 (2013). IEEE
19. Gemmeke, J.F., Ellis, D.P.W. List of iso 639-1 codes. https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes
20. Subtitle Guidelines. https://bbc.github.io/subtitle-guidelines
21. Auger, F., Flandrin, P., Lin, Y., McLaughlin, S., Meignen, S., Oberlin, T., Wu, H. Time-frequency reassignment and synchrosqueezing: An overview. IEEE Signal Process. Mag. **30**(6), 32–41 (2013). doi:10.1109/MSP.2013.2265316
22. Wang, L., Phapatanaburi, K., Oo, Z., Nakagawa, S., Iwahashi, M., Dang, J. Phase aware deep neural network for noise robust voice activity detection. In: 2017 IEEE International Conference on Multimedia and Expo, ICME 2017, Hong Kong, China, July 10-14, 2017, pp. 1087–1092 (2017). doi:10.1109/ICME.2017.8019414. https://doi.org/10.1109/ICME.2017.8019414
23. McCowan, I., Dean, D., McLaren, M., Vogt, R., Sridharan, S. The

delta-phase spectrum with application to voice activity detection and speaker recognition. IEEE Trans. Speech Audio Process. **19**(7), 2026–2038 (2011). doi:10.1109/TASL.2011.2109379
24. Chang, S., Li, B., Simko, G., Sainath, T.N., Tripathi, A., van den Oord, A., Vinyals, O. Temporal modeling using dilated convolution and gating for voice-activity-detection. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018, pp. 5549–5553 (2018). doi:10.1109/ICASSP.2018.8461921. https://doi.org/10.1109/ICASSP.2018.8461921
25. Lee, Y., Min, J., Han, D.K., Ko, H. Spectro-temporal attention-based voice activity detection. IEEE Signal Process. Lett. **27**, 131–135 (2020). doi:10.1109/LSP.2019.2959917
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I. Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
27. Zazo, R., Sainath, T.N., Simko, G., Parada, C. Feature learning with raw-waveform cldnns for voice activity detection. In: Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016, pp. 3668–3672 (2016). doi:10.21437/Interspeech.2016-268. https://doi.org/10.21437/Interspeech.2016-268
28. WebRTC VAD. https://webrtc.org/
29. Fonseca, E., Plakal, M., Font, F., Ellis, D.P.W., Serra, X. Audio tagging with noisy labels and minimal supervision. CoRR **abs/1906.02975** (2019). 1906.02975
30. Park, D.S., Chan, W., Zhang, Y., Chiu, C., Zoph, B., Cubuk, E.D., Le, Q.V. Specaugment: A simple data augmentation method for automatic speech recognition. In: Kubin, G., Kacic, Z. (eds.) Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019, pp. 2613–2617 (2019). doi:10.21437/Interspeech.2019-2680. https://doi.org/10.21437/Interspeech.2019-2680
31. Xie, S., Girshick, R.B., Dollár, P., Tu, Z., He, K. Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pp. 5987–5995
32. Phan, H., Koch, P., Katzberg, F., Maaß, M., Mazur, R., Mertins, A. Audio scene classification with deep recurrent neural networks. In: Lacerda, F. (ed.) Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017, pp. 3043–3047 (2017). http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0101.html
33. Scarpiniti, M., Comminiello, D., Uncini, A., Lee, Y. Deep recurrent neural networks for audio classification in construction sites. In: 28th European Signal Processing Conference, EUSIPCO 2020, Amsterdam, Netherlands, January 18-21, 2021, pp. 810–814 (2020). doi:10.23919/Eusipco47968.2020.9287802. https://doi.org/10.23919/Eusipco47968.2020.9287802
34. Can, D., Martinez, V.R., Papadopoulos, P., Narayanan, S.S. Pykaldi: A python wrapper for kaldi. In: Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference On (2018). IEEE
35. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K. The kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (2011). IEEE Catalog No.: CFP11SRW-USB
36. Bredin, H., Yin, R., Coria, J.M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., Gill, M.-P. pyannote.audio: neural building blocks for speaker diarization. In: ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing, Barcelona, Spain (2020)