Computer Speech and Language 74 (2022) 101349

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Taris: An online speech recognition framework with sequence to sequence neural networks for both audio-only and audio-visual speech

George Sterpu *, Naomi Harte

Sigmedia Lab, ADAPT Centre, Department of Electronic and Electrical Engineering, School of Engineering, Trinity College Dublin, Ireland

ARTICLE INFO

Keywords:
Online speech recognition
Audio-visual speech integration
Learning to count words
Multimodal speech processing
Speech recognition

ABSTRACT

It is widely accepted that the visual modality of speech provides complementary information to the speech recognition task, and many models have been introduced in order to make good use of the visual channel. This article develops Taris, a fully differentiable neural network model capable of decoding both audio-only and audio-visual speech in real time. We achieve this by connecting our previously proposed models AV Align and Taris, which are both end-to-end differentiable approaches to audio-visual speech integration and online speech recognition respectively. We evaluate AV Taris under the same conditions as AV Align and Taris on one of the largest publicly available audio-visual speech datasets, LRS2. Our results show that AV Taris is superior to the audio-only variant of Taris, demonstrating the utility of the visual modality to speech recognition within the real time decoding framework defined by Taris. Compared to an equivalent Transformer-based AV Align model that takes advantage of full sentences without meeting the real-time requirement, we report an absolute degradation of approximately 3% with AV Taris. As opposed to the more popular alternative for online speech recognition, namely the RNN Transducer, Taris offers a greatly simplified fully differentiable training pipeline. We speculate that AV Taris has the potential to popularise the adoption of Audio-Visual Speech Recognition (AVSR) technology and overcome the inherent limitations of the audio modality in less optimal listening conditions.[1]

1. Introduction

The next frontiers in computer speech technology include the capacity to enable natural conversations between humans and computers. An essential requirement from such technology is the automatic recognition of the spoken words with minimum latency. A fast and accurate decoding of speech by computers enables a more interactive communication with humans, creating the possibility for acknowledgements or interruptions to plead for clarifications. For example, we may be at a museum asking for directions, place an order at a restaurant, or check-in at a hotel lobby. In all these situations, it is important to guarantee a short response time to allow an interactive and natural conversation.

The remarkable progress in the age of deep learning has set the average word error rates on a par with the human level performance on relatively clean, structured conversations (Saon et al., 2017), and has enabled end-to-end automatic speech recognition, where the attention-based sequence to sequence (seq2seq) neural network can exceed the performance of the traditional

* Corresponding author.
E-mail address: sterpug@tcd.ie (G. Sterpu).

[1] Our code is publicly available at https://github.com/georgesterpu/Taris.

https://doi.org/10.1016/j.csl.2022.101349
Received 14 December 2020; Received in revised form 8 August 2021; Accepted 3 January 2022
Available online 19 January 2022

approaches (Chiu et al., 2018). Despite this, a distinct unresolved issue is reducing the latency from full utterances down to a few words. The seq2seq model conditions every target unit on the full unsegmented audio sentence, being predicated on the principle that a decoder drives the soft segmentation of the input during training. Because the first output token can only be emitted once the entire input sequence has been encoded, this sentence-level, or offline conditioning, is a fundamental barrier in decoding speech in real-time, or online, with a seq2seq network. It has been shown that, once convergence is reached, there are predominantly local relationships between the output tokens and the audio representations in speech (Chorowski et al., 2015; Chan et al., 2016). Therefore, potentially incurring no loss in accuracy, an explicit local conditioning of the outputs on the inputs would break the offline limitation and reduce the algorithmic latency. The new challenge is to learn robust associations between input and output subsequences which stand for the same linguistic concepts. We argue this is a necessary inductive bias in speech recognition, as the task specification sets no limit on maximum sequence lengths, and the truncation of long sentences is already performed during the collection of speech datasets.

Humans learn to simultaneously segment and recognise fluent speech from the earliest stages of life (Jusczyk and Aslin, 1995). Cairns et al. (1994) describe the relationship between speech segmentation and recognition as a chicken-and-egg problem: segmenting units with meaning (e.g. words) from continuous speech posits the recognition of the unit, but the recognition of a unit presumes its *a priori* segmentation. Unlike in text, there are no clear markers of the boundaries between the spoken words. Moreover, many words in a vocabulary represent the prefix of another longer word, such as *car* in *carbon*. Luce (1986) estimated that, prior to listening to the last phone of a word from a lexicon of 20,000 words in the English language, less than 50% of the most frequent words (up to 5 phonemes in length) are phonemically unique. Their study concluded that many words cannot be recognised in fluent speech until the initial segment of the following word is identified. Furthermore, Saffran et al. (1996) describe words as clusters of syllables characterised by a higher *transitional probability* between the syllables within the word than of the syllable pairs occurring at boundary between two consecutive words. Johnson and Jusczyk (2001) explain that we integrate a set of acoustic, phonetic, prosodic, and statistical cues in order to segment words in fluent speech. A natural pathway is to investigate whether explicitly introducing the ability to segment speech into *word* units with a neural network represents a reasonable inductive bias to address the challenge decoding speech in real-time. This approach would take advantage of the monotonicity of speech, allows the network focus on local properties, and removes the offline conditioning.

In our prior work (Sterpu et al., 2021), we investigated the research question of whether neural networks can learn to segment a spoken utterance by learning *to count* the words therein. Different from the way humans acquire a language, we suspected that there is a strong relationship between learning the ability to count spoken words, and the ability to segment words in fluent speech. We used this conjecture to develop Taris,[2], an online speech recognition system that estimates word boundaries, and subsequently eagerly decodes clusters of segments into words. Different from the related models, the network responsible for the analysis of the boundary cues in Taris is trained by minimising the difference between a word count estimated from the audio features and the ground truth value inferred from the text labels. For completeness reasons, we will review Taris in Section 3.

In more acoustically demanding situations involving distant microphones, overlapped speech, background noise, or natural dialogue structures, the speech recognition error rate stretches well past the human level performance (Barker et al., 2018; Watanabe et al., 2020). The visual modality of speech carries the potential to partially overcome these challenges and contribute to the sub-tasks of speaker diarisation, voice activity detection, and the recovery of the place of articulation, and can compensate for up to 15 dB of noise on average, as shown by the early studies of Sumby and Pollack (1954), Macleod and Summerfield (1987), Summerfield (1987). In this article, our research question extends the analysis of the segmentation approach in Taris to multiple modalities. We want to examine if words can be counted more reliably from the audio-visual representations of speech, than from audio alone. Binnie et al. (1974) find that the visual modality of speech provides cues regarding the place of articulation. Furthermore, Summerfield (1987) suggests that the visual modality may play a role in speech segmentation and voice activity detection. These properties are particularly important in noisy environments with overlapped speech, where we expect a superior speech recognition performance when integrating cues from multiple sources. Mitchel and Weiss (2010) found that the visual modality of speech does not improve the ability to segment an artificial language through statistical learning. They suspected that, under clean speech conditions, the boundary cues extracted from the audio modality were sufficiently robust and made the visual cues unnecessary. Later, Mitchel and Weiss (2014) investigated the role of visual cues for speech segmentation by creating an artificial language with limited statistical cues for the detection of word boundaries. They found the visual speech cues from talking faces to be beneficial to speech segmentation. They also ensured that the prosody cues inferred from vision did not contribute to segmentation by restricting the head nodding of the talking actors. The initial results of Tan and Burnham (2019) may be indicative of the fact that infants pay more attention to the mouth area when exposed to audio-visual speech stimuli. Taken together, this represents a strong motivation to investigate the potential of the visual modality to improve the robustness of Taris at decoding online.

To this end, we introduce *AV Taris*, a multimodal Transformer-based system for online speech recognition that extends Taris and operates on audio-visual speech inputs. We make use of our previously proposed method AV Align (Sterpu et al., 2018; Sterpu et al., 2020a) for fusing the two speech modalities. To enable the real time processing of the input sequences, we restrict the cross-modal alignment operation in AV Align to a window of fixed size. Subsequently, the network responsible with the prediction of the word count in Taris now has access to fused audio-visual speech cues instead of purely auditory ones. We find that AV Taris achieves a slightly higher error rate than the equivalent offline Audio-Visual Align Transformer model when trained and evaluated under the

[2] The name *Taris* echoes the misuse of the strong–weak syllable stress rule when learning to segment words by infants exposed to the phrase *the guitar is* discussed in Jusczyk et al. (1999).

same conditions, but is considerably superior to an audio-only model. Our result supports the hypothesis that the visual modality plays a significant role in the segmentation of a spoken utterance. Furthermore, the design of AV Taris preserves the property of Taris to decode speech in real time, on audio-visual inputs. We describe AV Taris in Section 3.3.

The rest of the article is structured as follows. In Section 2 we position Taris and AV Taris relative to other approaches in online speech recognition and audio-visual fusion. Section 3 introduces Taris and its audio-visual extension AV Taris. Section 4 presents our experiments with AV Taris on an audio-visual large vocabulary continuous speech dataset. In Section 5 we discuss the main findings and retrospect to the appropriateness of counting words to online decoding.

2. Related work

Two underlying problems modelled by AV Taris are the online decoding of speech and audio-visual speech integration. We will discuss below the key related work in these two areas.

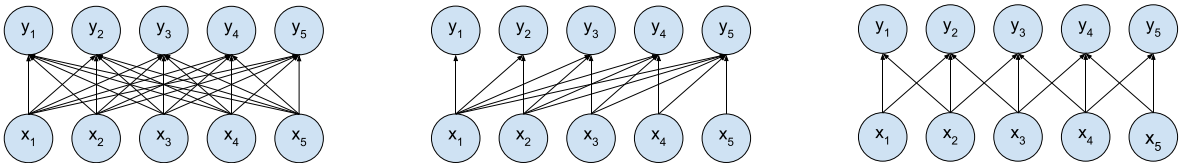
2.1. Online decoding

Real time, or online speech recognition is already possible with traditional speech models, or with neural network models based on the RNN Transducer (RNN-T) architecture of Graves (2012). The former models, including hidden Markov models with GMM/DNN class conditionals, despite generalising well on relatively small datasets for contemporary standards, are known to display diminishing returns for an increasing amount of speech data compared to their neural counterparts. As for the latter, Prabhavalkar et al. (2017) note in their comparative study that the inference in the RNN Transducer has the potential to be performed in a frame-synchronous, hence real time mode, if coupled with a unidirectional encoder. Their work only investigated bidirectional encoders to allow a more fair comparison to the attention model. More recently, the RNN-T model has been used in a practical setting of real time decoding. For example, Sainath et al. (2020) show that the RNN-T is comparable in latency and accuracy with a conventional model for only a fraction of the size. However, a shortcoming of RNN-T is its inference complexity, where two separate modules, the *Prediction* and *Transcription* networks, dynamically alternate their turns depending on the current output label being either a blank or non-blank token. Wang et al. (2019) analyse the shortcomings of the RNN-T, finding that its dynamic programming training algorithm marginalises over a large number of alignment paths including many unreasonable ones, and report training difficulties. Furthermore, Battenberg et al. (2017) note that bridging the modelling assumptions between the RNN-T and attention models, particularly by equipping attention models with the monotonicity constraints of the RNN-T, is a promising avenue. These attention models, commonly known as sequence to sequence (seq2seq) or encoder–decoder architectures (Cho et al., 2014; Sutskever et al., 2014; Forcada and Ñeco, 1997; Kalchbrenner and Blunsom, 2013), address the shortcomings of the RNN-T, but introduce a different computation paradigm that no longer enables online speech recognition. Such models were initially proposed in Machine Translation, and have not been fully optimised to the structure of the speech signal. Consequently, our aim in this work is to adapt sequence to sequence models for the online processing of speech utterances.

There have been several attempts to enable online decoding with sequence to sequence architectures. An early example is represented by the Triggered Attention strategy (TA) of Moritz et al. (2019). TA relies on the CTC decoder of Graves et al. (2006) to track changes in the frame-level CTC predictions, thereby informing an attention-based decoder of the boundaries between acoustic units, and allowing to limit the range of attention. Raffel et al. (2017) and Chiu and Raffel (2018) develop the concept of monotonic attention, which scans the entries in the attention memory from left to right in order to decide upon unit boundaries. Taris, our approach presented in Section 3 re-frames the online decoding problem for seq2seq networks such that it no longer requires a dynamic programming algorithm as in the examples above. Additional very specific related work will be presented in Section 3.7, since it is opportune to introduce Taris beforehand.

2.2. Audio-visual speech integration

Multimodal fusion of auditory and visual speech with a machine has been studied for almost four decades. Stork and Hennecke (1996) classify the initial attempts into early/feature fusion, intermediate fusion, and late/decision fusion. Robert-Ribes et al. (1996) and Schwartz et al. (1998) further discuss two additional categories, namely dominant modality recoding and motor/amodal recoding. While the discussion on the optimal fusion strategy for audio-visual speech has not yet settled to date, deep learning approaches went one step further in reducing the modelling assumptions by adopting a hybrid fusion model. As seen for example in Petridis and Pantic (2016), Makino et al. (2019), Petridis et al. (2018a), one simple approach involves transforming each modality into independent abstract representations, a concept similar to amodal recoding, then concatenating the two time-synchronous representations into a fused audio-visual representation, similar to early fusion. One notable work departing from this basic model is the Watch, Listen, Attend, and Spell (WLAS) network of Chung et al. (2017), which replaced the direct time-based fusion with a contextual fusion relying on an attention mechanism. More precisely, for each decoded output character, WLAS computes an auditory and a visual context vector that are both aligned with the decoder's current state, and then proceeds to combine the resulting contexts and states into a single multimodal representation. This approach presents the advantage that the two input modalities do not necessarily need to be time-synchronous. As a consequence, WLAS avoids a classic dilemma in which the well established audio and visual applications use different optimal sampling rates, typically 100 time–frequency audio vectors per second and approximately 30 video frames per second. Sterpu et al. (2018) challenge the design of Chung et al. (2017) regarding the place of integration, arguing that a decoder-based fusion under-exploits the low-level interactions between the auditory and visual modalities.



(a) Full connectivity (b) Causal connectivity (c) Windowed connectivity with one context frame to left and right

Fig. 1. Sequence-level conditioning strategies commonly used by neural network architectures in speech processing. For higher generality, x_i and y_i denote elements of generic input and output sequences, and will be specialised in this section as auditory and visual inputs and their corresponding internal representations.

In contrast with WLAS, Sterpu et al. (2018) introduce the AV Align architecture, where an attention mechanism is used to align each audio representation with every visual representation, thereby extracting a contextualised visual cue that is used to update and complement each audio frame. Later work of Sterpu et al. (2020a) shows that, when trained under very similar conditions on approximately 30 h of audio-visual recordings, AV Align has a clear edge over WLAS, particularly in noisy auditory conditions, whereas both strategies still outperform an audio-only model in terms of average error rates. One limitation of both strategies consists in aligning internal states over the full duration of a pre-segmented utterance, which precludes online decoding. This aspect motivated us to update the integration strategy of AV Align, and we will discuss it in Sections 3.2 and 3.3.

3. Taris

We begin this section by first reviewing the underlying approach Taris presented in Sterpu et al. (2021). Taris was originally designed to process the auditory modality of speech. In this work we will refer to it as *Audio Taris*, and we will discuss it in Section 3.1. We will extend this model to additionally take advantage of the visual modality of speech, introducing *AV Taris* in Section 3.3. Since the multimodal extension is a windowed variant of our previously proposed method *AV Align* (Sterpu et al., 2018), we will briefly summarise this method in Section 3.2. When referring to the overall approach based on word counting and segment-based real-time decoding, independent of the audio or audio-visual nature of the speech cues, we will only use the term *Taris*.

3.1. Audio Taris

The audio variant of Taris takes as input a variable length sequence of audio vectors $A = [a_1, a_2, \dots, a_N]$ and applies the Encoder stack of the Transformer model defined in Vaswani et al. (2017). A typical Transformer encoder has the sequence connectivity pattern illustrated in Fig. 1(a), where each representation in a higher order layer is conditioned on the entire sequence of representations from the previous layer. Since we aim to control the latency of the encoding process for speech signals, Audio Taris limits this conditioning to a fixed window centred on the sequence element with controlled look-back e_{LB} and look-ahead e_{LA} frames. Such an example for one look-ahead and look-back frames, respectively is displayed in Fig. 1(c). Technically, we achieve the restraint of the range of the attention operation in the self-attention Encoder of the Transformer by masking those attention weights outside the allowed range with zeros. The Encode operation below denotes the masked variant of the original operation in the Transformer, and produces the audio representations $O^A = [o_1^A, o_2^A, \dots, o_N^A]$:

$$O^A = \text{Encode}(A, e_{LB}, e_{LA}) \tag{1}$$

To obtain a soft, differentiable estimate of the word count from the encoder representations O^A, we start by applying a sigmoidal gating unit on each encoder output o_i^A to obtain a scalar score for each frame:

$$\alpha_i = \text{sigmoid}(o_i^A W_G + b_G) \tag{2}$$

where $\text{sigmoid}(x) = \frac{1}{1+\exp(-x)}$, $W_G \in \mathbb{R}^{h\times 1}$, $b_G \in \mathbb{R}^1$. Here, h stands for the number of neurons defining the size of the hidden state of our model.

We assign to every single input frame i a segment index $\hat{\imath o}_i$ by taking the *cumulative sum* of α and applying the *floor* function on the output:

$$\hat{\imath o}_i = \left\lfloor \sum_{j=1}^{i} \alpha_j \right\rfloor \tag{3}$$

Namely, the first predicted segment is delimited by a cumulative sum of α between 0 and 1, the second segment by the same quantity between 1 and 2, and so on.

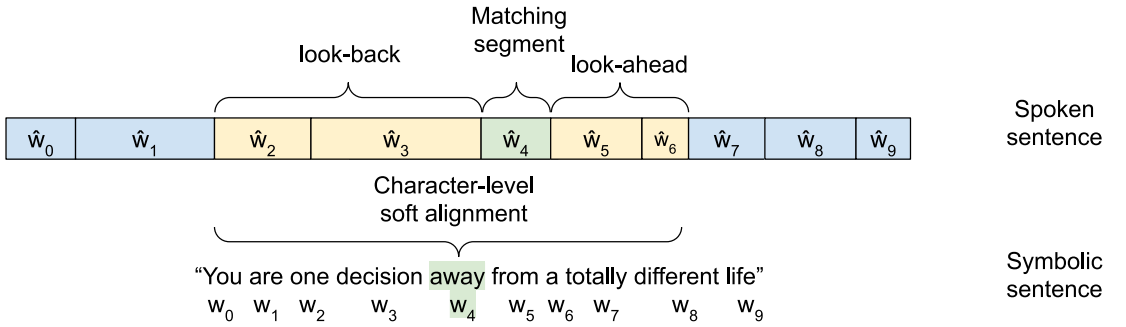


Fig. 2. The core idea behind Taris. A sub-network in the audio encoder incrementally assigns to each audio frame i a segment index $\hat{w}_i$. The Word Loss objective from Eq. (11) incentivises the estimation of a number of segments equal to the number of words in the training sample. Consequently, the decoder–encoder alignment can be now restricted between the characters of a given word and it corresponding audio segment, while also allowing a configurable look-back and look-ahead for the better modelling of short-term dependencies. In this example, when spelling the letters of the 4th word *away*, the decoder is allowed to attend to the audio frames that were assigned a segment index between 2 and 6 (illustration of $e_{LB} = e_{LA} = 2$).

During training, the Decoder stack receives the labelled grapheme sequence $Y = \{y_1, y_2, \dots, y_L\}$, made of English letters and the unique word delimiter SPACE. We assign to every grapheme k a word index w_k by leveraging the SPACE tokens in the labelled sequence:

$$w_k = \sum_{j=1}^{k} (y_j == \text{SPACE}) \tag{4}$$

Thus, whereas symbolic segmentation of speech uses a unique SPACE token to separate words, acoustic segmentation flags word boundaries by tracking the frame locations where the partial sum of the word counting signal α_i passes to the next integer value.

We modify the decoder–encoder connectivity of the Attention layer of Vaswani et al. (2017) to allow our decoder to perform soft-alignment over a *dynamic* window of segments estimated by the encoder. More precisely, we only allow those connections for which the following condition is met:

$$F = \widehat{W}_{ik} \leq (W_{ik} + d_{LA}) \textbf{ and } \widehat{W}_{ik} \geq (W_{ik} - d_{LB}) \tag{5}$$

In Eq. (5), d_{LA} and d_{LB} denote the number of segments the decoder is allowed to look-ahead and look-back respectively. The W and $\widehat{W}$ matrices are obtained from the w and $\hat{w}$ arrays by applying the tile operation, which repeats one sequence for a number of times equal to the length of the other one.

$$\widehat{W}_{ik} = \hat{w}_k \tag{6}$$

$$W_{ik} = w_k \tag{7}$$

$\forall i \in [1, N], \forall k \in [1, L]$. For example, if we assume a 4-word sentence with $w = [000111222333]$ and $\hat{w} = [0123]$, tiling generates two matrices W and $\widehat{W}$ of the same shape 12×4 by replicating the rows of w 4 times and the columns of $\hat{w}$ 12 times (and transposing).[3] We then extend the association between the indices of these matrices to support *segment* look-back and look-ahead. More generally, F is a 2D matrix $\in \mathbb{R}^{N \times L}$ that defines the admissible connections between any decoder timestep and any encoder timestep, acting as a bias on the decoder–encoder attention. Setting F as a matrix of ones recovers the original Transformer model. The extension to 3D tensors that include the batch dimension is straightforward, offering Taris efficient minibatch training and inference. Note that the look-back and look-ahead parameters of the encoder and the decoder denote different units. Whereas e_{LB} and e_{LA} designate a number of input frames, d_{LB} and d_{LA} represent clusters of frames that we call *segments*. The separation between *segments* and *words* is made on purpose to avoid making a strong claim regarding the nature of the speech segments at the end of system training. The overall process of matching audio segments $\hat{W}$ with the corresponding words W is illustrated in Fig. 2.

The decoder implements a traditional character level auto-regressive language model that predicts the next grapheme in the sequence, conditioned on all the previous characters and the dynamic audio context vector c_k:

$$c_k = \text{Attention}(\texttt{keys} = O^A, \texttt{query} = o^D_{k-1}, \texttt{mask} = F) \tag{8}$$

$$o^D_k = \text{Decode}(Y, c_k) \tag{9}$$

$$p_k \equiv P(y_k | c_k, y_{1:k-1}) = \text{softmax}(o^D_k W_\eta + b_\eta) \tag{10}$$

[3] A commonly used implementation of this operation is *numpy.tile*.

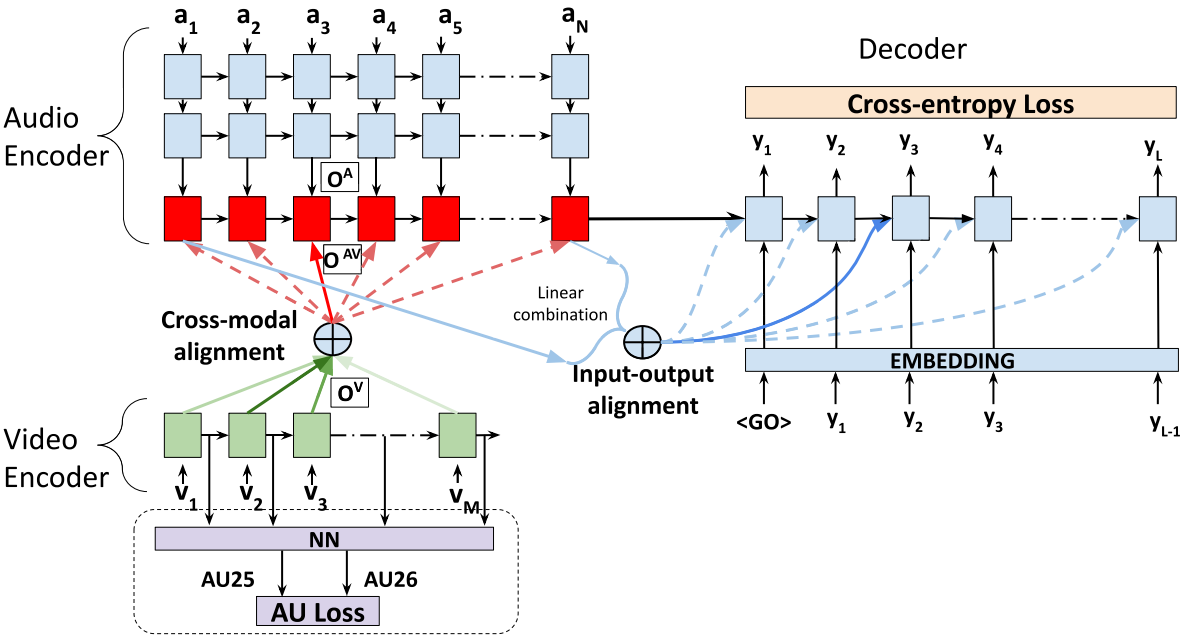


Fig. 3. The full AV Align strategy of audio-visual speech integration within the sequence to sequence framework for the task of speech recognition. Note that the two auditory and visual encoder stacks may be implemented with any neural transformation such as recurrent neural networks, Transformers, or convolutional layers (the diagram displays the former example).

where $W_\eta \in \mathbb{R}^{h \times \eta}, b_\eta \in \mathbb{R}^\eta$. In Eq. (10), η is the alphabet size of 28 tokens representing the 26 English letters, space, and apostrophe. We measure the difference between the estimated word sum $\hat{w_N} = \sum_{i=1}^{N} \alpha_i$ and the true word count $|w| = \sum_{k=1}^{L} (y_k == \text{SPACE})$ as:

$$\text{Word Loss} = (|w| - \hat{w_N})^2 \tag{11}$$

We define the training loss as:

$$\text{CE Loss} = \frac{1}{L} \sum_k -y_k \log(p_k) \tag{12}$$

$$\text{Loss} = \text{CE Loss} + \lambda \text{ Word Loss} \tag{13}$$

In all our experiments we used a scale factor $\lambda = 0.01$ found empirically. The self-attention connections of the Decoder are causal, describing an auto-regressive process.

3.2. AV align

Since we are interested in a design that learns to count words from audio-visual speech cues as opposed to audio-only ones seen in the previous section, we need a multimodal fusion strategy that integrates the two streams at an early stage. There are two reasonable choices for the integration step. First, there is direct feature concatenation, which expects a similar sampling rate of the two speech modalities to enable the sequence-level concatenation operation. Second, there is our previously proposed method AV Align (Sterpu et al., 2018; Sterpu et al., 2020a), that applies an attention mechanism from the audio modality to the visual one, and subsequently obtains time-aligned fused audio-visual speech representations based on the dot product correlation score. To bypass the equal sampling rate limitation of direct feature concatenation, in this work we will focus on AV Align as our multimodal speech integration strategy. We illustrate the full diagram of AV Align in Fig. 3.

Given a variable length acoustic sentence $A = [a_1, a_2, \dots, a_N]$ and its corresponding visual track $V = [v_1, v_2, \dots, v_M]$ of length $M \neq N$, two separate stream encoders (e.g. Recurrent/Convolutional Neural Networks, Transformer) project the sequences onto higher order abstract representations $O^A = [o_1^A, o_2^A, \dots, o_N^A]$ and $O^V = [o_1^V, o_2^V, \dots, o_M^V]$ as follows:

$$O^A = \text{Encode}(A) \tag{14}$$

$$O^V = \text{Encode}(V) \tag{15}$$

AV Align obtains a fused sequence $O^{AV} = [o_1^{AV}, o_2^{AV}, \dots, o_N^{AV}]$ where the audio representations o_i^A are fused with their contextualised visual representations c_i^V using an attention mechanism:

$$\alpha_{ij} = \text{softmax}_i(o_i^{AT} \cdot o_j^V) \tag{16}$$

$$c_i^V = \sum_{j=1}^{M} \alpha_{ij} \cdot o_j^V \tag{17}$$

where $\text{softmax}_i(X) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$.

In order to fuse O_i^A with the corresponding c_i^V, our initial work (Sterpu et al., 2020a) used a learnable transformation taking as input the concatenation of the two vector representations:

$$o_i^{AV} = W_{AV}[o_i^A; c_i^V] + b_{AV} \tag{18}$$

where $W_{AV} \in \mathbb{R}^{\text{h}\times 2\text{h}}, b_{AV} \in \mathbb{R}^{\text{h}}$.

Later, with a Transformer encoder replacing the recurrent one, we chose in Sterpu et al. (2020b) a simpler fusion step without additional learnable parameters:

$$o_i^{AV} = o_i^A + c_i^V \tag{19}$$

Since the Audio-Visual Transformer model in Sterpu et al. (2020b) took advantage of the visual modality of speech at a similar rate as the LSTM variant in Sterpu et al. (2020a) compared to an equivalent audio-only model, in this work we will adopt the simpler fusion approach in Eq. (19) with no learnable parameters. This will keep the overall number of parameters to a minimum level, therefore increasing the chance of correct generalisation on a relatively small dataset.

Sterpu et al. (2020a) also show that a sequence to sequence architecture using the AV Align modality integration strategy does not learn useful visual representations when randomly initialising the parameters of the audio-visual model. The most likely outcome under visually challenging conditions was that the trained model learns to ignore the visual modality, achieving a decoding accuracy comparable with a single modality audio-only system. This suggests that stronger regularisation may be needed for the audio-visual speech recognition task. Related works of Chung et al. (2017), Afouras et al. (2018), or (Petridis et al., 2018b) pre-train the visual front-end on a separate visual word classification task, whereas (Tao and Busso, 2021) use an multitask learning structure combining speech recognition with voice activity detection from audio-visual cues. In this work we use the Action Unit (AU) auxiliary objective proposed in Sterpu et al. (2020a). Specifically, we use the OpenFace toolkit of Baltrusaitis et al. (2018) to obtain the target intensities of two AUs, *Lips Part* (AU25) and *Jaw Drop* (AU26), and introduce a linear layer estimating the two AUs from the visual representations:

$$\widehat{AU}_{25,26}(j) = \text{sigmoid}(W_{AU}\, o_j^V + b_{AU}) \tag{20}$$

where $W_{AU} \in \mathbb{R}^{2\times n}, b_{AU} \in \mathbb{R}^2$. The auxiliary objective measures the mean squared difference between the target and predicted AUs:

$$AU\ \ Loss = \frac{1}{M}\sum_{j=1}^{M}(AU_{25,26}(j) - \widehat{AU}_{25,26}(j))^2 \tag{21}$$

AV Align lacks the specification of the temporal limits of integration, which is an inherent limitation of the attention mechanism. Specifically, any fused representation o_i^{AV} is conditioned on each visual representation o_j^V for any j from 1 to M. Consequently, this scheme would no longer allow us to control the encoding latency that was achieved with Taris. In the next section we will address this challenge and will propose an audio-visual extension of Taris based on a windowed version of AV Align.

3.3. Audio-visual Taris

The visual modality of speech does not contain sufficient linguistic information to allow the prediction of word boundaries. As a result, we cannot use the same counting strategy as with the audio modality in order to segment visual speech. Learning to segment the audio modality was necessary because the auditory and symbolic modalities of speech exist on different timescales, and we found the concept of words as the linking element between them. However, the audio and video modalities share the same time axis and can be integrated more easily, by only taking into account the different sampling rates. Having prior knowledge of the natural asynchrony between auditory and visual speech allows us to set an upper limit on the audio-visual integration window.

We describe an audiovisual extension of Taris. Considering the previously defined audio and visual sequences A and V, and their encoded representations O^A and O^V respectively, we define a symmetrical integration window of length $2B+1$ centred on a visual frame index j and apply a *constrained* cross-modal alignment between modalities:

$$c_i^V = Attention\ (o_i^A, o^V{}_{j-B\ :\ j+B}) \tag{22}$$

$$j = \left\lfloor (i+1)\frac{N}{M} \right\rfloor - 1 \tag{23}$$

For any audio frame i, the index j is calculated as the nearest time-aligned video frame, e.g. audio frame 50 corresponds to the video frame 25 when the audio has twice the sampling rate of the video (i.e. $N = 2M$). Compared to c_i^V in Eq. (5) from the offline multimodal architecture used in our prior work (Sterpu et al., 2020a), here the alignment is performed within a window of $2B+1$ visual frames, which is only a fraction of the full length M of the visual sequence. Consequently, an audio representation only depends on temporally local video representations, preserving the eager decoding property of Taris. The audio and visual representations are integrated as following:

$$o^{AV} = C^V + O^A \tag{24}$$

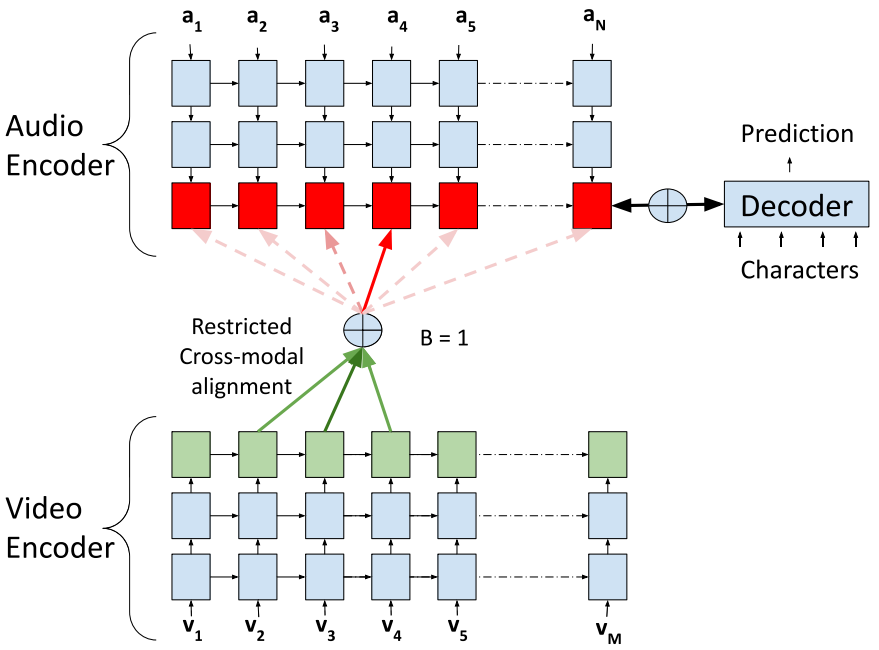


Fig. 4. Zoom on the audio-visual integration process of AV Taris restricted to a window of 2B+1 frames (B=1 in this case). The 4th audio frame a_4 is time-synchronous with the 3rd video frame v_3, thus cross-modal attention is performed over v_2, v_3, v_4.

Fig. 4 highlights the restricted alignment process described by Eqs. (22)–(23).

To complete the online audio-visual design, we only need to predict the gating signal α_i from the fused representations o_{AV} instead of the audio ones seen in Eq. (2):

$$\alpha_i = \text{sigmoid}(o_i^{AV} W_G + b_G) \tag{25}$$

This strategy allows us to investigate if AV Taris can learn to count words in fluent speech better from the fused audiovisual representations instead of the audio ones alone.

3.4. Latency analysis

Calculating the delay between the first audio frame timestamp and the first output unit is non trivial and depends on several factors.

First, the encoding look-ahead and look-back parameters e_{LA} and e_{LB} define the receptive field in audio frames of a learnt audio representation o_i^A. The absolute encoding delay is a function of e_{LA}, the number of layers in the Transformer network, and the audio frame duration. We provide in Section 4.4 a more precise measurement of the encoding latency.

Second, each audio representation o_i^A requires up to B future frames from the visual modality for the multimodal re-contextualisation step. The larger value between the receptive field of the self-attention operation in the last layer of the audio encoder and B will give the encoding latency of the multimodal system.

Third, the decoding look-ahead parameter d_{LA} defines the number of audio segments required to start decoding the first grapheme in an output word unit. Although this number of segments is constant, the frame length of d_{LA} segments is dynamic and context dependent. In other terms, the first grapheme in the next word can be decoded once the cumulative sum of α_i becomes greater or equal to d_{LA}.

Finally, we note that an alternative decoding approach in Taris is to gradually increase the segment look-ahead from 0 to d_{LA}, and consequently to provide up to $d_{LA}+1$ updates for the same word. This can be more practical when an immediate, less accurate transcription is needed, accepting that it is subject to corrections depending on the future context.

A quantitative analysis of the system latency, succinctly regarded as the earliest timestamp to print a word after the occurrence of the last sound in the word, is a challenging task that would require a substantial extension of our work. One difficulty is given by the internal structure of Taris, which does not offer any guarantees that the acoustic segments described by w_i match the word units described by $\hat{w}$. Eventually, this metric could be calculated in a straightforward way when there are no word-level insertions or deletions, thus allowing a one-to-one correspondence between segments and words. Consequently, in this work we relax the above definition of latency, and will display two related quantities instead. One is the word counting loss from Eq. (11), which will tell how well can Taris approximate the true word count. Furthermore, we will contrast the histograms of the segment lengths with the actual durations of the words at the corpus level, which can potentially reveal issues around under- or over-segmentation.

3.5. Complexity analysis

When operating on a single modality, Taris requires a negligible overhead in parameters and operations over the original Transformer. The only extra parameters are given by the W_G and b_G variables in Eqs. (2), which amount for $h+1$ scalars in the total model size. Eqs. (2)–(5) describe the additional operators mainly consisting of a matrix vector multiplication followed by a sigmoid activation for every audio frame, and the update of a scalar cumulative sum. Since attention masking is already performed by the original Transformer to take into account the true input length in a minibatch, Eq. (8) does not represent an overhead, and the only additional operation needed at each decoded timestep is the computation of the segment mask F in Eq. (5). In training, this mask is computed only once per batch, since we have access to the full output sequence and know the positions of all the SPACE tokens in advance. The mask F is directly applicable on the tensor product performed by the Transformer architecture between the queries and keys by adding a large negative value outside the mask before applying the softmax operation. This grants Taris a highly efficient computation strategy that integrates with the Transformer.

On the other hand, AV Taris additionally requires the application of a Convolutional Neural Network (CNN) on the input images, and an encoder of the same nature as the one used for the audio representations. A thorough discussion on the computational complexity of various CNN architectures is out of the scope of this article. Nevertheless, the overall strategy we propose is invariant of the specific CNN and sequence encoders options, which may be optimised with respect to the requirements of the intended application.

3.6. Considered alternatives

Attention-based sequence to sequence models learn an explicit alignment between the output tokens and the speech frames. We initially considered leveraging the alignments corresponding to each SPACE token of a pre-trained offline model, and using this as a supervision signal to train the gate α_i from Eq. (2). However, visually inspecting these alignments revealed that there is no clear delimitation between the spoken words in English, with the softmax weights not being skewed towards a low number of frames. This observation is in line with our intuition that the SPACE token rarely corresponds to a short pause in English speech, and instead has a more analytic role which demands its inference from the acoustic differences between several words, or from the intrinsic language model in the decoder. In effect, such alignment information would only represent a crude approximation of the true boundaries between the spoken words, and it would be a very noisy supervision signal for our gate α_i to learn.

In Taris, we can choose to constrain the output of the gating unit α_i in Eq. (2) to follow a specific distribution. For example, Hou et al. (2020) train their gating unit to follow a Bernoulli distribution, making values very close to 0, or very close to 1, more likely. During our initial experiments with a scaled sigmoid function (i.e. $1/(1+exp(-kx)), k>1$), we noticed that this unit does not typically have values close to the extremities of the range, and achieves a slightly higher word counting loss than standard unscaled sigmoid. We speculate that the gating unit learns to *accumulate* cues at the sub-word level in order to solve the word counting task, and an eventual binary output behaviour may only be feasible when coupled with a recurrent process to keep track of an internal state, which in turn would complicate the design. In our work, α_i is predicted directly from the hidden state o_{A_i} with a feed-forward neural network.

As an alternative to the sigmoid activation we also considered the hyperbolic tangent function, which has an output range between $(-1..1)$. The negative output values have the potential to enable a broader range of word counting strategies, such as assigning higher confidence scores and eventually correcting them later based on future evidence. With the sigmoid activation, the system does not have the opportunity to make corrections and has to adopt a more defensive approach. On the other hand, a sufficiently large receptive field may reduce the need for such corrections. In our initial experiments we did not see a significantly improved word loss with the $tanh$ activation, and, since Taris is a relatively new model, we decided to apply the law of parsimony and maintain the $sigmoid$ until empirical evidence demands otherwise.

3.7. Comparison to related work

Dong et al. (2020) categorise end-to-end speech recognition models into label-synchronous and frame-synchronous models. The former refers to models that derive the contextual acoustic units from the soft alignment with the state of an auto-regressive decoder receiving grapheme labels as inputs. In contrast, the latter derive acoustic labels directly from the audio representations by removing the decoder, and thereby do not model the conditional dependence between the labels.

Taris is more closely related to the label-synchronous class, as it maintains a soft attention mechanism between the decoder and the encoder. However, Taris derives a segmentation signal directly from the audio representations, and the soft alignment is only allowed within a well defined dynamic window. This contrasts with the model proposed by Dong and Xu (2020), which also predicts a normalised weight per frame, but uses these weights directly once they sum up to approximately 1.0 to linearly combine the corresponding audio representations into a single state from which the segment label is estimated. An approach similar to the one of Dong and Xu (2020) was previously introduced in Li et al. (2019), however they only test the method on Mandarin speech. Li et al. (2019) anticipate problems on languages such as English with less clear boundaries between linguistic units and complex orthographies.

Taris is more closely related to the approach of Hou et al. (2020) performing segment level attention. However, Hou et al. (2020) take a different approach to train the boundary detection unit by sampling from a Bernoulli distribution, which makes the model non-differentiable, and resort to policy gradients. Experimentally, they find that the cumulative sum of the boundary unit requires a

dynamic threshold ranging from 0.2 to 0.55 for optimal decoding performance. This suggests that the approach we take with Taris not enforcing a specific distribution on the output of the gating unit, and only requiring the total sum to be close to the word count, is likely enabling the learning of a more flexible counting mechanism. The sigmoidal unit in Taris does not enforce the notion of a hard boundary, but instead we design the decoder to analyse a limited acoustic range covered by the cumulative sum of the gating unit.

Wang et al. (2020) present a related method that uses of a dynamic encoding context to update the audio representations, as does Taris. Our methods differ in the strategy used to detect the boundaries of the dynamic context. Wang et al. (2020) propose a *scout network* to scan for possible word boundaries. Their approach is similar to Hou et al. (2020) or (Li et al., 2019), as it aims to classify each audio frame as either a word boundary or non-boundary. Instead of training the scout network with a single objective function as in the prior work, Wang et al. (2020) see it as a supervised task and make use of forced alignment to provide frame-level word boundary labels. This creates the requirement of a pre-trained forced aligner model, which increases the complexity of the training pipeline. It also represents a possible source of noise when the forced aligner produces noisy or uncertain labels. Instead, our method Taris does not require external labels to train the gating unit, beyond the trivial word count of the considered sentence. Moreover, Taris does not enforce the notion of a word boundary in its encoder, and allows the gating unit to take any intermediate value in the [0..1) range. It is only the decoder of Taris that constructs segments by analysing the cumulative sum of the gate output. We believe these are important advantages of Taris that allow a very simple training. On the other hand, without boundary supervision, Taris does not guarantee that it can learn word boundaries. Instead, the reliability of its boundary predictions is highly coupled with the sentence diversity of the training corpus.

In the space of audio-visual speech recognition, another model that is capable of decoding online is the work of Makino et al. (2019). Their model is based on the RNN-T, and uses a simpler direct feature fusion strategy for audio-visual integration. They report tailoring the audio feature extraction strategy to match the video frame rate by shifting the spectral analysis window with variable increments. Instead, AV Taris uses the seq2seq architecture as a back-end for sequence modelling, avoiding the training difficulties specific to the RNN-T. Furthermore, we leverage the cross-modal fusion strategy AV Align multimodal feature integration, which decouples the sampling rates of the two input streams.

4. Experiments and results

We conduct our experiments on the unconstrained audio-visual English speech dataset LRS2 (BBC and University of Oxford, 2017). The main training set of LRS2, used in this work, contains 45,839 spoken sentences, and the test set contains 1243 sentences. We use the same dataset partitioning, the same audio features, and the same strategy for corruption with additive cafeteria noise at a SNR of 10db, 0db, and −5 dB as in Sterpu et al. (2021) to enable a direct comparison between Audio Taris and AV Taris. Our implementation of Taris forks the official Transformer model in TensorFlow 2 (The TensorFlow Model Garden, 2020).

4.1. Input pre-processing

Our system takes auditory and visual input concurrently. The **audio** input is the raw waveform signal of an entire sentence. The **visual** stream consists of video frame sequences, centred on the speaker's face, which correspond to the audio track. We use the OpenFace toolkit of Baltrusaitis et al. (2018) to detect and align the faces, then we crop around the lip region using a static bounding box extracted from the bottom half of the aligned face.[4] Complete details of the pre-processing of each stream now follow.

Audio input. The audio waveforms in LRS2 have a sampling rate of 16,000 Hz. The audio signals are additively mixed with cafeteria noise at different Signal to Noise Ratios (SNR) as explained in Section 4.2. We compute the log magnitude spectrogram of the input, choosing a frame length of 25 ms with 10 ms stride and 1024 frequency bins for the Short-time Fourier Transform (STFT), and a frequency range from 80 Hz to 11,025 Hz with 30 bins for the mel scale warp. We stack the features of 8 consecutive STFT frames into a larger window, leading to an audio feature vector a_i of size 240, and we shift this window right by 3 frames, thus attaining an overlap of 5 frames between consecutive audio windows.

Visual input. We down-sample the 3-channel RGB images of the lip regions to 36 × 36 pixels. A ResNet CNN (He et al., 2016) processes the images to produce a 256-dimensional feature vector v_j. The details of the architecture are presented in Table 1.

4.2. Training procedure

The acoustic modality is corrupted with only *Cafeteria* noise, as this noise type was found the most challenging in Sterpu et al. (2018), and the noise source did not influence the conclusions. We train our systems in four stages, first on clean speech, then with a Signal to Noise Ratio (SNR) of 10db, 0db and finally −5db. Each time we increment the noise level we also copy the model parameters rather than train from scratch, speeding up the system's convergence. We train our models on LRS2 for a total of 120 epochs at an initial learning rate of 0.001, decayed to 0.0001 after 100 epochs, on each noise level. The systems are evaluated at the end of the 120 iterations. The training time of AV Taris is approximately 450 s for a single epoch of LRS2 on an Nvidia Titan XP GPU.

[4] Precise coordinates can be found in the publicly available code linked on the first page of the article.

Table 1
CNN architecture. All convolutions use 3 × 3 kernels, except the final one. The residual Block (He et al., 2016) is in its *full preactivation* variant.

Layer	Operation	Output shape
0	Rescale [−1 ...+1]	36 × 36 × **3**
1	Conv	36 × 36 × **8**
2–3	Res block	36 × 36 × **8**
4–5	Res block	18 × 18 × **16**
6–7	Res block	9 × 9 × **32**
8–9	Res block	5 × 5 × **64**
10	Conv 5 × 5	1 × 1 × **256**

For the reasons explained in Sterpu et al. (2021), we do not use the common End-of-Sentence (EOS) token to pad our target sequences when training our online models. While remaining auto-regressive, the decoding process for each sentence in the training minibatch stops after predicting the expected number of words from the audio input. In English and other analytic languages (in contrast with synthetic agglutinating languages), this can be trivially derived from the count of blank space tokens in the decoding history.

4.3. Neural network details

Our models use 6 layers for two Encoders (audio and video) and Decoder stacks of the Transformer, with a hidden model size $d_{model} \equiv h = 256$, a filter size $d_{ff} = 256$, one attention head, and 0.1 dropout on all attention weights and feedforward activations. The audio-visual models occupy 36 MB on disk, and are considerably smaller than the typical size of state-of-the-art models used in benchmarks. We chose this model size so we could train it on a single GPU of 12 GB of memory with a minibatch size of 32. We presume that a larger model may bring a similar level of improvement if we wanted to pursue a better absolute accuracy. This would come at the cost of slower, more expensive training iterations, and eventually of stronger regularisation requirements when considering the amount of training data.

In the following sections, we will refer to the models relying on the Taris decoding/inference strategy as *online* models, whereas the baseline Transformer models, with or without the auxiliary word counting objective, will be referred to as *offline* models, since they use a full decoder–encoder attention connectivity.

4.4. Analysis of the receptive field

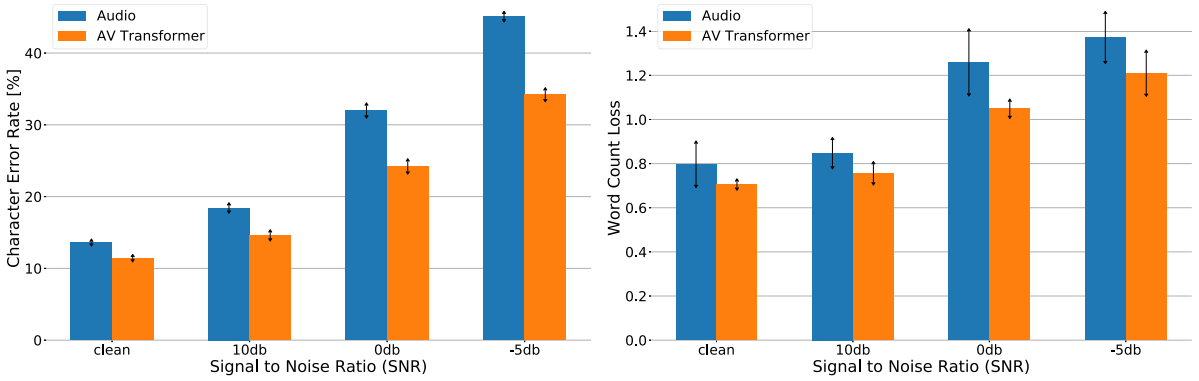
As we described our data pre-processing setup in Section 4.1, one audio frame is obtained by stacking 8 STFT frames taken over 25 ms windows with 10 ms strides. Each new audio frame includes the previous 5 STFT frames, so the additional non-overlapping information is represented by 3 STFT frames. In greater detail, the first audio frame achieves an effective range from 0 ms to 95 ms. The second audio frame starts at 30 ms going up to 125 ms, followed by the third frame from 60 ms to 155 ms, and so on.

The first layer in our Transformer encoder network has a receptive field in frames controlled by the e_{LB} and e_{LA} parameters. We preserve the same mask throughout the entire Transformer stack. This means that the superior layers can access a broader receptive field with respect with the audio input. A representation at position k in the Transformer layer l is then indirectly conditioned on the audio input up to the position $k + l \cdot e_{LA}$. We leave the fine tuning of this connectivity design for latency optimisation as future work.

4.5. Learning to count words in audio-visual speech

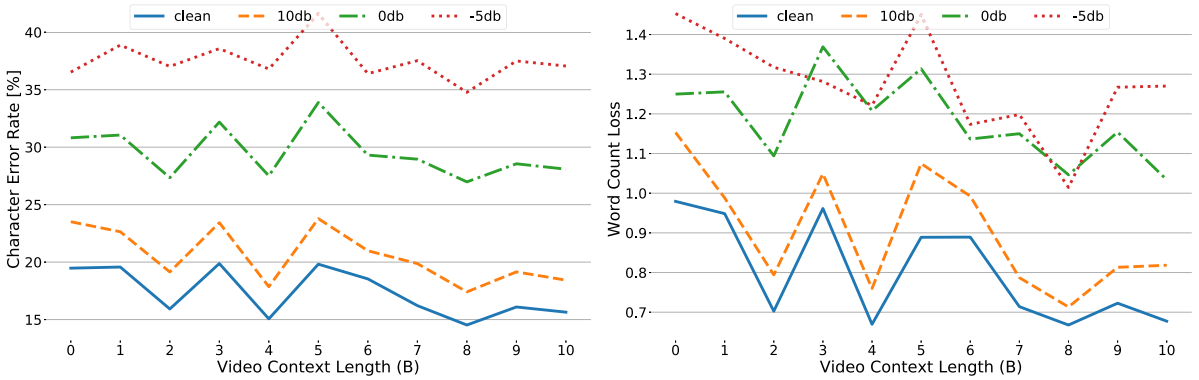
We are interested in studying if the word count in fluent speech can be estimated with a higher accuracy from audio-visual cues than from the audio modality alone. In Sterpu et al. (2021) we saw that the encoding look-ahead length does not have a major influence on either the word counting error or the character error rate. Therefore, in this experiment we limit our analysis to counting words from audio-visual representations with the offline models having infinite context available. We train Audio and Audio-Visual Transformer models on LRS2 and repeat the experiment for five different random initialisations. We plot the average Character Error Rate and the Word count loss of the two systems in Fig. 5. The arrows indicate the standard deviation across the five trials. To help the visual network learn good representations, the Audio-Visual Transformer uses the auxiliary Action Unit loss described in Sterpu et al. (2020a). Note that the Transformer-based models used in this experiment take advantage of the entire input sequence when updating each representation, and cannot be used in an online setting.

From Fig. 5(b) it can be seen that the average word count loss of the Audio-Visual Transformer is slightly lower than the one of the Audio model, while the recognition accuracy shown in Fig. 5(a) stays approximately the same as in our prior work (Sterpu et al., 2020b), where we did not use the Word Loss. This aspect suggests that the visual cues may be informative of word boundaries in fluent speech, although it is difficult to draw conclusions regarding the statistical significance of this result from only 5 trials.



(a) Character Error Rate (b) Word Count Loss

Fig. 5. Evaluation of the offline Audio and the Audio-Visual Transformer on LRS2 with the word counting loss enabled.



(a) Character Error Rate (b) Word Counting Loss

Fig. 6. Evaluation of AV Taris on LRS2 when varying the size of the symmetrical window used for the soft-selection of the visual representation aligned with each audio representation.

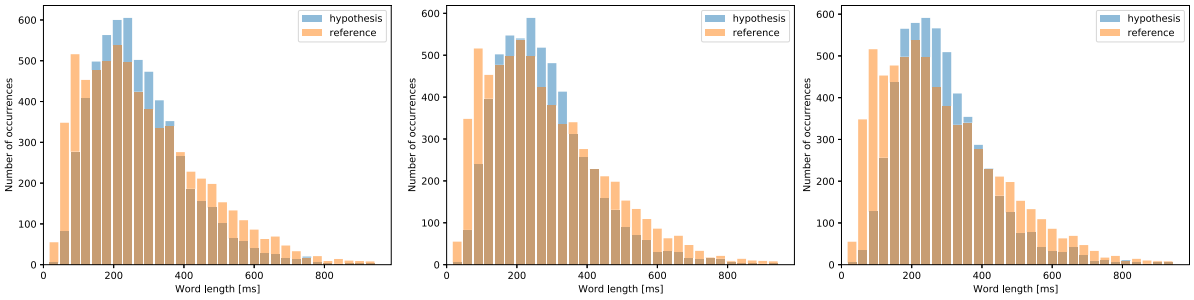
4.6. Online audio-visual decoding

The decoder in our previous experiment had access to the entire encoder memory. For our online models in this section we opt for an encoder look-ahead e_{LA} of 11 frames and infinite look-back $e_{LB} = \infty$, since we showed in Sterpu et al. (2021) that there are diminishing gains beyond this threshold. In that work we have also demonstrated that Audio Taris can leverage the gating signal α to limit the dynamic range of decoder–encoder attention and still match the error rate of the offline Transformer. We now investigate how the audio-visual extension of Taris compares to the offline Audio-Visual Transformer.

In Section 4.5 we have seen that an Audio-Visual Transformer with infinite look-back and look-ahead encoding context achieves a slightly lower word counting loss than an Audio-only counterpart. Therefore, learning to count from the fused audio-visual representation does not degrade the word counting accuracy. The system may additionally take advantage of the visual modality to further improve its counting estimate.

We evaluate the AV Taris model for an increasing length of the cross-modal attention window, controlled by the length parameter B. The window length is defined as $len(w) = 2B+1$, as it extends symmetrically in both directions. When $B = 0$ the system is similar to a down-sampled version of feature fusion that bypasses the requirement to have identical sampling rates for both modalities. For $B > 0$, the window is extended with B frames to the left and to the right respectively. The results obtained with AV Taris are displayed in Fig. 6.

Our best AV Taris systems are obtained when B is equal to 2, 4, and greater than 7. Given the difficulty of evaluating a neural network with respect to all possible source of variations such as weight initialisation, example shuffling or random dropout, which remains an open research question (see Dror et al. (2019) for a thorough analysis), it is impractical to perform a proper significance test for the systems displayed in Fig. 6. One notable trend is that the error rates stabilise with the increasing length of the video attention window. Beyond 7 future video frames there are diminishing returns in decoding accuracy, although it may be possible to further restrict the context to only 2 video frames and still achieve comparable results.



(a) Audio Taris clean (b) AV Taris clean (c) AV Taris -5db

Fig. 7. Segmentation length distribution (in milliseconds) of Audio-only and Audio-visual Taris compared to the reference provided by the Montreal forced aligner.

We observe an absolute difference of approximately 3% at all the noise levels between AV Taris and the offline AV Align Transformer model shown in Fig. 5(a). AV Taris is still superior to the audio-only models both in their offline and online variants at higher levels of noise. However, in low noise conditions up to 10db, the degradation of AV Taris over the AV Align Transformer considerably diminishes the benefit of cross-modal alignment over audio-only modelling. Nevertheless, the contribution of the visual modality in clean speech may span beyond the average Character Error Rate metric used in this work. This remains a topic for further investigation.

It is important to note that the error rates presented above are considerably higher than the best results obtained on LRS2 by Petridis et al. (2018b) and Yu et al. (2021), i.e. character error rates of 3.6% and 2.4% respectively on clean speech. In particular, Petridis et al. use direct feature integration and add a secondary objective function for sequence alignment, while also pre-training the video front-end on a separate visual word classification task and on the much larger pre-training partition of LRS2 (approximately 200 h of recordings), and finally interpolate the model output with an external language model. Similarly, Yu et al. use a pre-trained video front-end, then fine-tune their Transformer-based model on larger pre-train partitions of LRS2 and the newer LRS3, while they also leverage an external language model. As expected, the amount of training data has a crucial role on the final system performance, making the discussion more relevant from the perspective of developing higher quality products than allowing a fair comparison between different modelling strategies. Instead, the aim of the present article is to introduce a new simple technique for sequence to sequence models capable of decoding offline, and different from the existing approaches to the online decoding of speech discussed in Section 2.1. We leave the training and evaluation of AV Taris on increasing amounts of data as future work.

4.7. Segment length histograms

As in Sterpu et al. (2021), we want to analyse the typical lengths of the segments that AV Taris discovers. Although it would be much stronger to investigate the correlation between segments and words more explicitly, beyond duration, we appreciate there are several drawbacks for doing so. First, Taris does not explicitly optimise the correct segmentation of words from a spoken utterance, but this is only a by-product of the Character Error Rate objective. Furthermore, since Taris allows left and right context when decoding a word, this diminishes the necessity for precise word boundaries. Instead, ensuring a reasonable duration for the segments Taris discovers is essential to low latency decoding, which is the main objective of our work.

In Fig. 7 we plot the histograms of the segment lengths, revealing the audio context used during the decoding process. To compute the length of a segment, we first detect those timestamps where the cumulative sum of the gating signal α_i passed to an integer value. Next, we compute the difference between any two consecutive crossing points. We then transform these values representing frame counts to milliseconds by multiplying them with the same amount that was used to shift the audio feature window to the right when pre-processing the audio data, namely 30 ms. Finally, we compute the histograms of all the segment lengths from all the sentences in the test set of LRS2. Fig. 7(a) displays the histogram corresponding to Audio Taris on clean speech, whereas Figs. 7(b)–7(c) show the histograms of AV Taris on clean and noisy speech respectively. In orange, we overlay the histogram of the ground truth/reference word lengths estimated with the Montreal forced aligner tool of McAuliffe et al. (2017).

In all cases, we notice that the word counting strategy adopted in AV Taris leads to a reasonable approximation of the word length distribution estimated with the forced aligner. At this level of coarseness, it is hard to appreciate the differences in segmentation between Audio Taris and AV Taris on clean speech. The only noticeable difference is between clean and −5db speech, where the mean of the distribution shifts to the right in the latter case. This suggests that the AV Taris system has the tendency to create longer segments on noisy speech by fusing several shorter segments detected on clean speech. The overall effect would be an increased decoding latency on noisy speech, suggesting that the system requires more audio context before becoming sufficiently confident to initiate decoding.

5. Discussion and conclusion

We have proposed a multimodal extension of the Taris system as a fully differentiable solution to online audio-visual speech recognition. To accomplish this, we revised the cross-modal attention mechanism in AV Align by limiting the attention span to a fixed window of video representations centred on each audio frame. As a result, we have achieved an audio-visual speech recognition system that can decode online. Experimentally, we found that the accuracy of the audio-visual extension of Taris lags behind the offline Transformer-based AV Align system approximately by an absolute 3%. The offline model could exploit the entire utterance for both cross-modal alignment and decoding, explaining one possible source for the difference. We believe that the modelling assumptions in both AV Align and Taris are sufficiently general to transfer to other multimodal speech processing tasks.

Chiu et al. (2019) and Narayanan et al. (2019) report that the neural networks used in speech recognition struggle to generalise to sentences that are considerably longer than the ones seen in training. Instead, Taris only models the local relationships in speech, and is structurally unaffected by the sentence length at inference beyond the analysis window defined by the look-back and look-ahead parameters. The same property was seen in the Neural Transducer (NT) model of Jaitly et al. (2016), although Taris allows adaptive segments and simplified training by avoiding the problems introduced by the NT's end-of-block token. Similar to the NT model, Taris repeatedly applies a sequence to sequence model over consecutive audio windows. The NT model processes fixed length blocks, and does not need to *learn* a segmentation. Their mechanism increases the complexity of the optimisation algorithm. More precisely, the model introduces an additional end-of-block token in the output domain that needs to be emitted once per every audio window. This generates the problem of having to search for an optimal alignment in training between the longer sequence of predicted labels containing the additional token and the shorter ground truth sequence. Taris avoids this problem by not making use of end-of-block tokens. Instead, Taris analyses dynamic windows of speech centred on a word of interest. On the other hand, Taris does not guarantee the reliable segmentation of the spoken utterance into words. It only facilitates the compensation of eventual segmentation errors with a controllable number of look-back and look-ahead segments that the decoder is allowed to attend to. Studying the internal segmentation achieved by Taris remains a topic for future exploration.

Compared to alternative online models such as the RNN Transducer, Taris reduces the computational cost of training and the engineering cost of maintaining the hardware-specific software implementation of the RNN-T objective function. Additionally, it springs from the sequence to sequence model architecture that is currently outperforming alternative approaches. We believe that both the audio and the audio-visual variants of Taris represent a step forward for increasing the accessibility of audio-visual speech recognition technology, although they still require validation at a much larger scale than this article could afford. Considering the current limits of ASR technology to accurately decode highly unstructured and noisy speech as seen in the recent CHiME-5 challenge (Barker et al., 2018), we believe that the original contributions of this work will enliven the adoption of AVSR solutions.

An interesting behaviour of Taris concerns the handling of word contractions, such as *you're, that's, don't, it's, let's*, and others. In our work, we considered that written words are exclusively separated by spaces, as seen in Eq. (4). Unless there is a systematic error in the transcriptions, Taris has the potential to learn the acoustic differences between "you're" and "you are". The system maintains its own segment counter (the cumulative sum of α_i), and has sufficient freedom to decide which form to transcribe. When "you are" is more likely, then the fraction of α_i added to $\hat{w_N}$ may simply be one unit greater than when "you're" is preferred. Taris can also recover from potential errors since it uses a context window larger than a single segment. Depending on the intermediate scores α_i, the decoding of "are" in "you are" may then be conditioned on the acoustic representations corresponding to "you" and other adjacent segments. On the other hand, the general formulation of the word counting task in Taris may be problematic in the case of modelling silences. Since silences are generally not annotated in the human transcriptions, Taris implicitly includes all the audio frames not substantially modifying α_i to the adjacent segment. Consequently, the decoder performs a soft alignment even over those uninformative silence frames. Increasing the efficiency of this process represents a possible direction of improvement.

It is unlikely that humans learn to segment speech by counting words in full sentences. We are not offered the word count in a numeric format as a supervision signal. Why would it be appropriate to design a speech recognition system based on this aspect? We believe there are several reasons. First, this task would not be impossible for humans if it was formulated as a puzzle for finding patterns in a foreign language. Language acquisition in humans involves a long term process of teaching simpler, isolated words before gradually increasing the difficulty. These learning strategies have not fully matured in our machine learning technology. On the other hand, it is very common, and cheap, to produce a speech dataset annotated at the sentence level, without intermediate phone-level or word-level alignments. Therefore we are already asking computers to solve the speech recognition challenge differently from the way we learn a spoken language. We argue that learning to count words is a good compromise with respect to our existing technology and datasets when aiming to segment a spoken utterance.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to thank Christian Saam for the feedback offered on the manuscript. We are also grateful to the anonymous reviewers for their very constructive feedback.

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106 at the ADAPT SFI Research Centre at Trinity College Dublin, Ireland. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant # 13/RC/2106. Our work is supported by a Titan Xp GPU grant from NVIDIA, United States.

References

Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A., 2018. Deep audio-visual speech recognition. IEEE Trans. Pattern Anal. Mach. Intell. http://dx.doi.org/10.1109/TPAMI.2018.2889052, 1–1.

Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L., 2018. OpenFace 2.0: Facial Behavior analysis toolkit. In: 13th IEEE International Conference On Automatic Face Gesture Recognition. pp. 59–66. http://dx.doi.org/10.1109/FG.2018.00019.

Barker, J., Watanabe, S., Vincent, E., Trmal, J., 2018. The Fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In: Proc. Interspeech 2018. pp. 1561–1565. http://dx.doi.org/10.21437/Interspeech.2018-1768.

Battenberg, E., Chen, J., Child, R., Coates, A., Li, Y.G.Y., Liu, H., Satheesh, S., Sriram, A., Zhu, Z., 2017. Exploring neural transducers for end-to-end speech recognition. In: 2017 IEEE Automatic Speech Recognition And Understanding Workshop (ASRU). pp. 206–213. http://dx.doi.org/10.1109/ASRU.2017.8268937.

BBC and University of Oxford, 2017. The Oxford-BBC lip reading sentences 2 (LRS2) dataset. Online, http://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html. (Accessed 4 May 2020).

Binnie, C.A., Montgomery, A.A., Jackson, P.L., 1974. Auditory and visual contributions to the perception of consonants. J. Speech Hear. Res. 17 (4), 619–630. http://dx.doi.org/10.1044/jshr.1704.619.

Cairns, P., Shillcock, R., Chater, N., Levy, J., 1994. Lexical segmentation: The role of sequential statistics in supervised and un-supervised models, in: Proceedings Of The 16th Annual Conference Of The Cognitive Science Society, pp. 136–141.

Chan, W., Jaitly, N., Le, Q., Vinyals, O., 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: 2016 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP). pp. 4960–4964. http://dx.doi.org/10.1109/ICASSP.2016.7472621.

Chiu, C., Han, W., Zhang, Y., Pang, R., Kishchenko, S., Nguyen, P., Narayanan, A., Liao, H., Zhang, S., Kannan, A., Prabhavalkar, R., Chen, Z., Sainath, T., Wu, Y., 2019. A comparison of end-to-end models for long-form speech recognition. In: 2019 IEEE Automatic Speech Recognition And Understanding Workshop (ASRU). pp. 889–896. http://dx.doi.org/10.1109/ASRU46091.2019.9003854.

Chiu, C.-C., Raffel, C., 2018. Monotonic chunkwise attention. In: International Conference On Learning Representations. pp. 1–16, URL https://openreview.net/forum?id=Hko85plCW.

Chiu, C., Sainath, T.N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R.J., Rao, K., Gonina, E., Jaitly, N., Li, B., Chorowski, J., Bacchiani, M., 2018. State-of-the-Art Speech recognition with sequence-to-sequence models. In: 2018 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP). pp. 4774–4778. http://dx.doi.org/10.1109/ICASSP.2018.8462105.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings Of The 2014 Conference On Empirical Methods In Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp. 1724–1734. http://dx.doi.org/10.3115/v1/D14-1179.

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y., 2015. Attention-based models for speech recognition. In: Proceedings Of The 28th International Conference On Neural Information Processing Systems. MIT Press, pp. 577–585.

Chung, J.S., Senior, A., Vinyals, O., Zisserman, A., 2017. Lip reading sentences in the wild. In: 2017 IEEE Conference On Computer Vision And Pattern Recognition (CVPR). pp. 3444–3453. http://dx.doi.org/10.1109/CVPR.2017.367.

Dong, L., Xu, B., 2020. CIF: Continuous Integrate-and-fire for end-to-end speech recognition. In: ICASSP 2020 - 2020 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP). pp. 6079–6083. http://dx.doi.org/10.1109/ICASSP40776.2020.9054250.

Dong, L., Yi, C., Wang, J., Zhou, S., Xu, S., Jia, X., Xu, B., 2020. A comparison of label-synchronous and frame-synchronous end-to-end models for speech recognition. arXiv:2005.10113.

Dror, R., Shlomov, S., Reichart, R., 2019. Deep dominance - How to properly compare deep neural models. In: Proceedings Of The 57th Annual Meeting Of The Association For Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 2773–2785. http://dx.doi.org/10.18653/v1/P19-1266.

Forcada, M.L., Ñeco, R.P., 1997. Recursive hetero-associative memories for translation. In: Mira, J., Moreno-Díaz, R., Cabestany, J. (Eds.), Biological And Artificial Computation: From Neuroscience To Technology. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 453–462. http://dx.doi.org/10.1007/BFb0032504.

Graves, A., 2012. Sequence transduction with recurrent neural networks, in: ICML Representation Learning Worksop, pp. 1–8.

Graves, A., Fernández, S., Gomez, F., Schmidhuber, J., 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Proceedings Of The 23rd International Conference On Machine Learning. In: ICML '06, Association for Computing Machinery, New York, NY, USA, pp. 369–376. http://dx.doi.org/10.1145/1143844.1143891.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. In: ECCV 2016. Springer International, pp. 630–645. http://dx.doi.org/10.1007/978-3-319-46493-0_38.

Hou, J., Guo, W., Song, Y., Dai, L.-R., 2020. Segment boundary detection directed attention for online end-to-end speech recognition. EURASIP J. Audio Speech Music Process. 2020 (1), 3. http://dx.doi.org/10.1186/s13636-020-0170-z.

Jaitly, N., Le, Q.V., Vinyals, O., Sutskever, I., Sussillo, D., Bengio, S., 2016. An online sequence-to-sequence model using partial conditioning. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (Eds.), Advances In Neural Information Processing Systems 29. Curran Associates, Inc., pp. 5067–5075, URL http://papers.nips.cc/paper/6594-an-online-sequence-to-sequence-model-using-partial-conditioning.pdf.

Johnson, E.K., Jusczyk, P.W., 2001. Word segmentation by 8-Month-Olds: When speech cues count more than statistics. J. Mem. Lang. 44 (4), 548–567. http://dx.doi.org/10.1006/jmla.2000.2755.

Jusczyk, P., Aslin, R., 1995. Infants' detection of the sound patterns of words in fluent speech. Cogn. Psychol. 29 (1), 1–23. http://dx.doi.org/10.1006/cogp.1995.1010.

Jusczyk, P.W., Houston, D.M., Newsome, M., 1999. The beginnings of word segmentation in english-learning infants. Cogn. Psychol. 39 (3), 159–207. http://dx.doi.org/10.1006/cogp.1999.0716.

Kalchbrenner, N., Blunsom, P., 2013. Recurrent continuous translation models. In: Proceedings Of The 2013 Conference On Empirical Methods In Natural Language Processing. Association for Computational Linguistics, Seattle, Washington, USA, pp. 1700–1709, URL https://www.aclweb.org/anthology/D13-1176.
Li, M., Liu, M., Masanori, H., 2019. End-to-end speech recognition with adaptive computation steps. In: ICASSP 2019 - 2019 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP). pp. 6246–6250. http://dx.doi.org/10.1109/ICASSP.2019.8682500.
Luce, P.A., 1986. A computational analysis of uniqueness points in auditory word recognition. Percept. Psychophys. 39 (3), 155–158. http://dx.doi.org/10.3758/BF03212485.
Macleod, A., Summerfield, Q., 1987. Quantifying the contribution of vision to speech perception in noise. Br. J. Audiol. 21 (2), 131–141. http://dx.doi.org/10.3109/03005368709077786, PMID: 3594015.
Makino, T., Liao, H., Assael, Y., Shillingford, B., Garcia, B., Braga, O., Siohan, O., 2019. Recurrent neural network transducer for audio-visual speech recognition. In: 2019 IEEE Automatic Speech Recognition And Understanding Workshop (ASRU). pp. 905–912. http://dx.doi.org/10.1109/ASRU46091.2019.9004036.
McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M., 2017. Montreal forced aligner: Trainable text-speech alignment using Kaldi. In: Proc. Interspeech 2017. pp. 498–502. http://dx.doi.org/10.21437/Interspeech.2017-1386.
Mitchel, A.D., Weiss, D.J., 2010. What's in a face? Visual contributions to speech segmentation. Lang. Cogn. Process. 25 (4), 456–482. http://dx.doi.org/10.1080/01690960903209888.
Mitchel, A.D., Weiss, D.J., 2014. Visual speech segmentation: using facial cues to locate word boundaries in continuous speech. Lang. Cogn. Neurosci. 29 (7), 771–780. http://dx.doi.org/10.1080/01690965.2013.791703.
Moritz, N., Hori, T., Roux, J.L., 2019. Triggered attention for end-to-end speech recognition. In: ICASSP 2019 - 2019 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP). pp. 5666–5670. http://dx.doi.org/10.1109/ICASSP.2019.8683510.
Narayanan, A., Prabhavalkar, R., Chiu, C., Rybach, D., Sainath, T.N., Strohman, T., 2019. Recognizing long-form speech using streaming end-to-end models. In: 2019 IEEE Automatic Speech Recognition And Understanding Workshop (ASRU). pp. 920–927. http://dx.doi.org/10.1109/ASRU46091.2019.9003913.
Petridis, S., Pantic, M., 2016. Deep complementary bottleneck features for visual speech recognition. In: 2016 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP). pp. 2304–2308. http://dx.doi.org/10.1109/ICASSP.2016.7472088.
Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., Pantic, M., 2018a. End-to-end audiovisual speech recognition. In: 2018 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP). pp. 6548–6552. http://dx.doi.org/10.1109/ICASSP.2018.8461326.
Petridis, S., Stafylakis, T., Ma, P., Tzimiropoulos, G., Pantic, M., 2018b. Audio-visual speech recognition with a hybrid CTC/Attention architecture. In: 2018 IEEE Spoken Language Technology Workshop (SLT). pp. 513–520. http://dx.doi.org/10.1109/SLT.2018.8639643.
Prabhavalkar, R., Rao, K., Sainath, T.N., Li, B., Johnson, L., Jaitly, N., 2017. A comparison of sequence-to-sequence models for speech recognition. In: Proc. Interspeech 2017. pp. 939–943. http://dx.doi.org/10.21437/Interspeech.2017-233.
Raffel, C., Luong, M.-T., Liu, P.J., Weiss, R.J., Eck, D., 2017. Online and linear-time attention by enforcing monotonic alignments. In: Proceedings Of The 34th International Conference On Machine Learning - Volume 70. In: ICML'17, JMLR.org, pp. 2837–2846.
Robert-Ribes, J., Piquemal, M., Schwartz, J.-L., Escudier, P., 1996. Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition. In: Stork, D.G., Hennecke, M.E. (Eds.), Speechreading By Humans And Machines: Models, Systems, And Applications. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 193–210. http://dx.doi.org/10.1007/978-3-662-13015-5_14.
Saffran, J.R., Newport, E.L., Aslin, R.N., 1996. Word segmentation: The role of distributional cues. J. Mem. Lang. 35 (4), 606–621. http://dx.doi.org/10.1006/jmla.1996.0032.
Sainath, T.N., He, Y., Li, B., Narayanan, A., Pang, R., Bruguier, A., Chang, S., Li, W., Alvarez, R., Chen, Z., Chiu, C., Garcia, D., Gruenstein, A., Hu, K., Kannan, A., Liang, Q., McGraw, I., Peyser, C., Prabhavalkar, R., Pundak, G., Rybach, D., Shangguan, Y., Sheth, Y., Strohman, T., Visontai, M., Wu, Y., Zhang, Y., Zhao, D., 2020. A streaming on-device end-to-end model surpassing server-side conventional model quality and latency. In: ICASSP 2020 - 2020 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP). pp. 6059–6063. http://dx.doi.org/10.1109/ICASSP40776.2020.9054188.
Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.-L., Roomi, B., Hall, P., 2017. English conversational telephone speech recognition by humans and machines. In: Proc. Interspeech 2017. pp. 132–136. http://dx.doi.org/10.21437/Interspeech.2017-405.
Schwartz, J.-L., Robert-Ribes, J., Escudier, P., 1998. Ten years after summerfield: A taxonomy of models for audio–visual fusion in speech perception. In: Hearing By Eye II: Advances In The Psychology Of Speechreading And Auditory-Visual Speech. Psychology Press/Erlbaum (UK) Taylor & Francis, Hove, England, pp. 85–108.
Sterpu, G., Saam, C., Harte, N., 2018. Attention-based audio-visual fusion for robust automatic speech recognition. In: Proceedings Of The 20th ACM International Conference On Multimodal Interaction. In: ICMI '18, ACM, New York, NY, USA, pp. 111–115. http://dx.doi.org/10.1145/3242969.3243014.
Sterpu, G., Saam, C., Harte, N., 2020a. How to teach DNNs to pay attention to the visual modality in speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 1052–1064. http://dx.doi.org/10.1109/TASLP.2020.2980436.
Sterpu, G., Saam, C., Harte, N., 2020b. Should we hard-code the recurrence concept or learn it instead? exploring the transformer architecture for audio-visual speech recognition. In: Proc. Interspeech 2020. pp. 3506–3509. http://dx.doi.org/10.21437/Interspeech.2020-2480.
Sterpu, G., Saam, C., Harte, N., 2021. Learning to count words in fluent speech enables online speech recognition. In: 2021 IEEE Spoken Language Technology Workshop (SLT). pp. 38–45. http://dx.doi.org/10.1109/SLT48900.2021.9383563.
Stork, D.G., Hennecke, M.E., 1996. Speechreading: an overview of image processing, feature extraction, sensory integration and pattern recognition techniques, in: Proceedings Of The Second International Conference On Automatic Face And Gesture Recognition, pp. XVI–XXVI.
Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26 (2), 212–215. http://dx.doi.org/10.1121/1.1907309.
Summerfield, Q., 1987. Some preliminaries to a comprehensive account of audio-visual speech perception. In: Hearing By Eye: The Psychology Of Lip-Reading.. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, pp. 3–51.
Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: Proceedings Of The 27th International Conference On Neural Information Processing Systems - Volume 2. In: NIPS'14, MIT Press, Cambridge, MA, USA, pp. 3104–3112.
Tan, S.H.J., Burnham, D., 2019. Auditory-visual speech segmentation in infants. In: Proc. The 15th International Conference On Auditory-Visual Speech Processing. pp. 43–46. http://dx.doi.org/10.21437/AVSP.2019-9.
Tao, F., Busso, C., 2021. End-to-end audiovisual speech recognition system with multitask learning. IEEE Trans. Multimed. 23, 1–11. http://dx.doi.org/10.1109/TMM.2020.2975922.
The TensorFlow Model Garden, 2020. Transformer Translation Model. Online, https://github.com/tensorflow/models/tree/r2.1.0/official/nlp/transformer. (Accessed 25 May 2020).
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Advances In Neural Information Processing Systems 30. Curran Associates, Inc., pp. 5998–6008.
Wang, D., Wang, X., Lv, S., 2019. An overview of end-to-end automatic speech recognition. Symmetry 11 (8), http://dx.doi.org/10.3390/sym11081018.
Wang, C., Wu, Y., Lu, L., Liu, S., Li, J., Ye, G., Zhou, M., 2020. Low latency end-to-end streaming speech recognition with a scout network. In: Proc. Interspeech 2020. pp. 2112–2116. http://dx.doi.org/10.21437/Interspeech.2020-1292.
Watanabe, S., Mandel, M., Barker, J., Vincent, E., Arora, A., Chang, X., Khudanpur, S., Manohar, V., Povey, D., Raj, D., Snyder, D., Subramanian, A.S., Trmal, J., Yair, B.B., Boeddeker, C., Ni, Z., Fujita, Y., Horiguchi, S., Kanda, N., Yoshioka, T., Ryant, N., 2020. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. In: 6th International Workshop On Speech Processing In Everyday Environments (CHiME 2020). Barcelona, Spain, URL https://hal.inria.fr/hal-02546993.
Yu, W., Zeiler, S., Kolossa, D., 2021. Fusing information streams in end-to-end audio-visual speech recognition. In: ICASSP 2021 - 2021 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP). pp. 3430–3434. http://dx.doi.org/10.1109/ICASSP39728.2021.9414553.