ADAPT
Engaging Content
Engaging People

A World Leading SFI Research Centre

Science Foundation Ireland sfi For what's next

# Multimodal speech understanding

Prof. Naomi Harte
Trinity College Dublin, Ireland
7th December 2021

HOST INSTITUTION
Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

HOST INSTITUTION
DCU Ollscoil Chathair Bhaile Átha Cliath Dublin City University

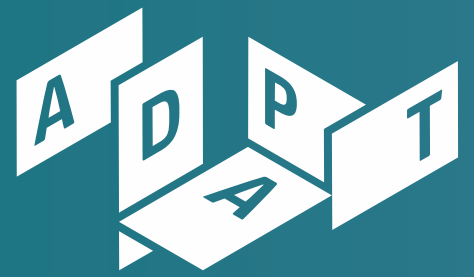
PARTNER INSTITUTIONS
TU Dublin
UCD University College Dublin An Coláiste Ollscoile, Baile Átha Cliath Ireland's Global University
MTU Ollscoil Teicneolaíochta na Mumhan Munster Technological University
AIT Connect & Discover
Maynooth University National University of Ireland Maynooth
OÉ Gaillimh NUI Galway

FUNDED BY:
Ireland's European Structural and Investment Funds Programmes 2014-2020 Co-funded by the Irish Government and the European Union
European Union European Regional Development Fund
Science Foundation Ireland sfi For what's next

# Today…

- Why multimodal?
- Two core aspects of speech
- Deep learning architectures
  - Audio visual speech recognition
  - Multimodal turn-taking prediction

# Speech Recognition today




**Engaging Content**
Engaging People

# (Audio-only) Speech Recognition

- Where is ASR is now?
- NIST Hub5 2000
- Switchboard
  - Strangers, 36 of 40 speakers in test are in train set
- CallHome
  - Conversational speech with familiarity

| | WER SWB | WER CH |
|---|---|---|
| Transcriber 1 raw | 6.1 | 8.7 |
| Transcriber 1 QC | 5.6 | 7.8 |
| Transcriber 2 raw | 5.3 | 6.9 |
| Transcriber 2 QC | **5.1** | **6.8** |
| Transcriber 3 raw | 5.7 | 8.0 |
| Transcriber 3 QC | 5.2 | 7.6 |
| Human WER from [1] | 5.9 | 11.3 |

Table 1: Word error rates on SWB and CH for human transcribers before and after quality checking contrasted with the human WER reported in [1].

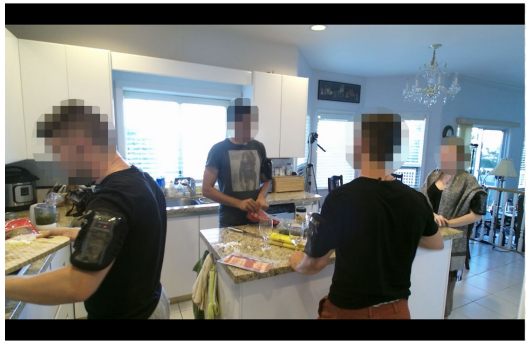
| | WER [%] | |
|---|---|---|
| | SWB | CH |
| n-gram | 6.7 | 12.1 |
| n-gram + model-M | 6.1 | 11.2 |
| n-gram + model-M + Word-LSTM | 5.6 | 10.4 |
| n-gram + model-M + Char-LSTM | 5.7 | 10.6 |
| n-gram + model-M + Word-LSTM-MTL | 5.6 | 10.3 |
| n-gram + model-M + Char-LSTM-MTL | 5.6 | 10.4 |
| n-gram + model-M + Word-DCC | 5.8 | 10.8 |
| n-gram + model-M + 4 LSTMs + DCC | **5.5** | **10.3** |

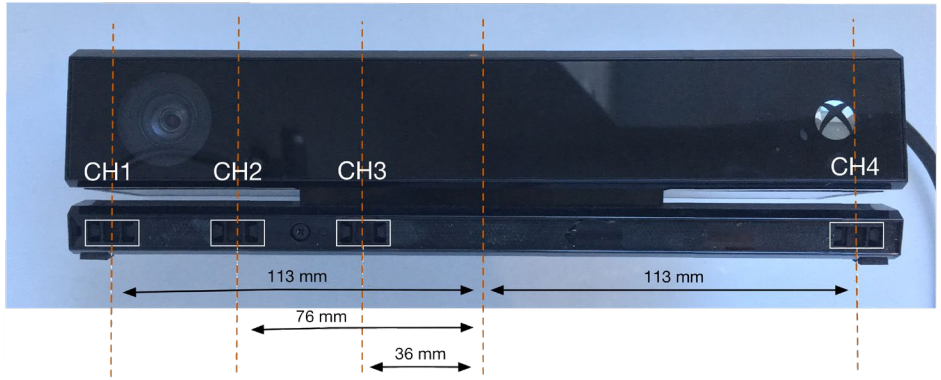
Table 8: *WER on SWB and CH with various LM configurations.*

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.-L., Roomi, B., and Hall, P. (2017). *English conversational telephone speech recognition by humans and machines*. In Proc. Interspeech 2017, pages 132–136.

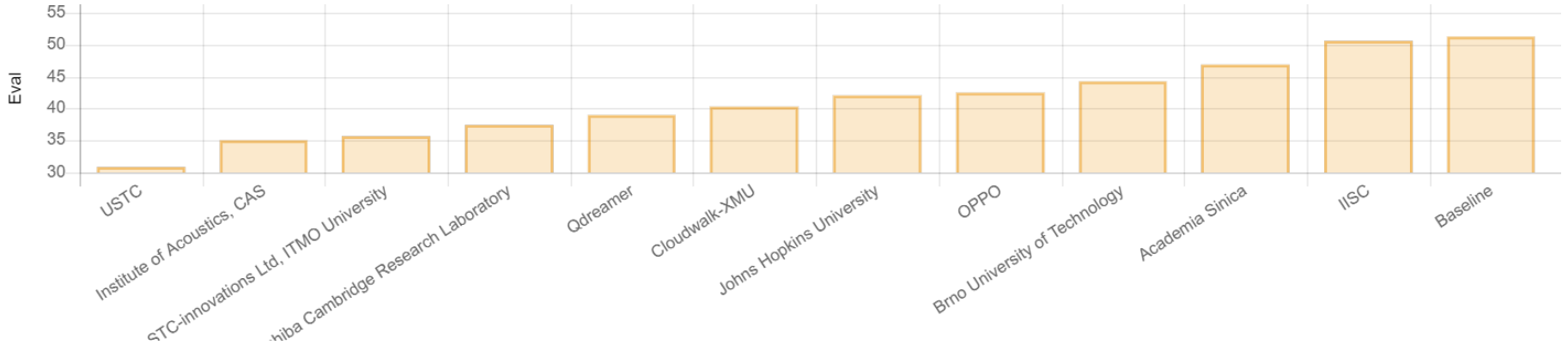
# What about noise?

- 6th CHiME Speech Separation and Recognition Challenge (CHiME-6)
- Distant microphone conversational speech recognition
- Everyday home environments
  - 20 parties each recorded in a different home
  - Binaural microphones (to synchronise only) worn by each participant (4 participants per session), and by 6 microphone arrays with 4 microphones each

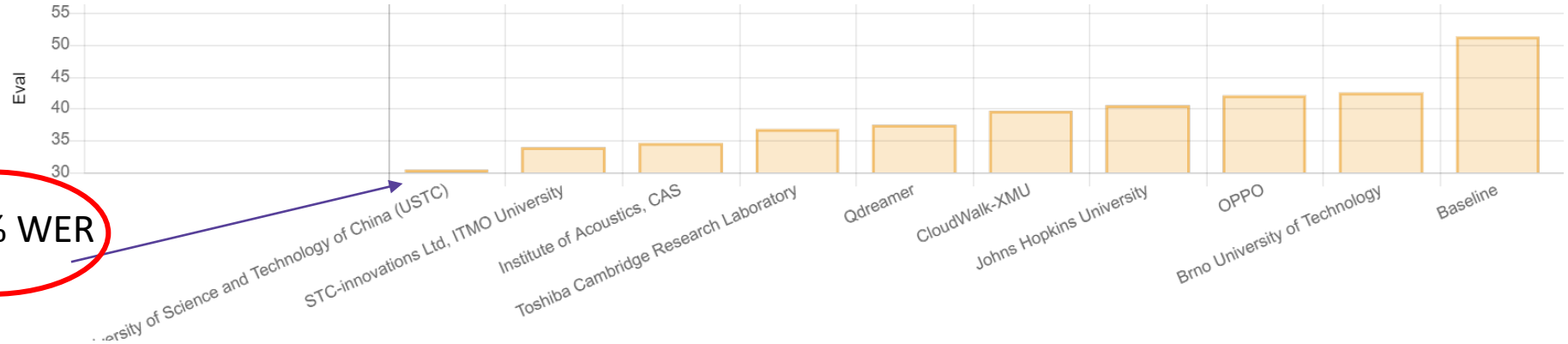




# ASR - Track 1




Constrained Language model



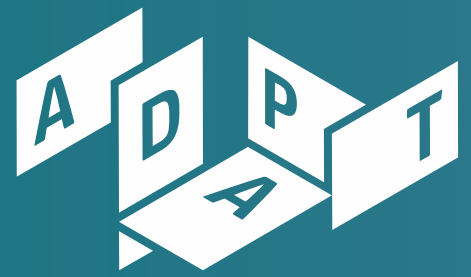

Unconstrained Language model

30.51% WER

# ASR limitations

- Noise
- Overlap
- Non-native speakers
- Children, elderly
- Atypical
- Not just ASR

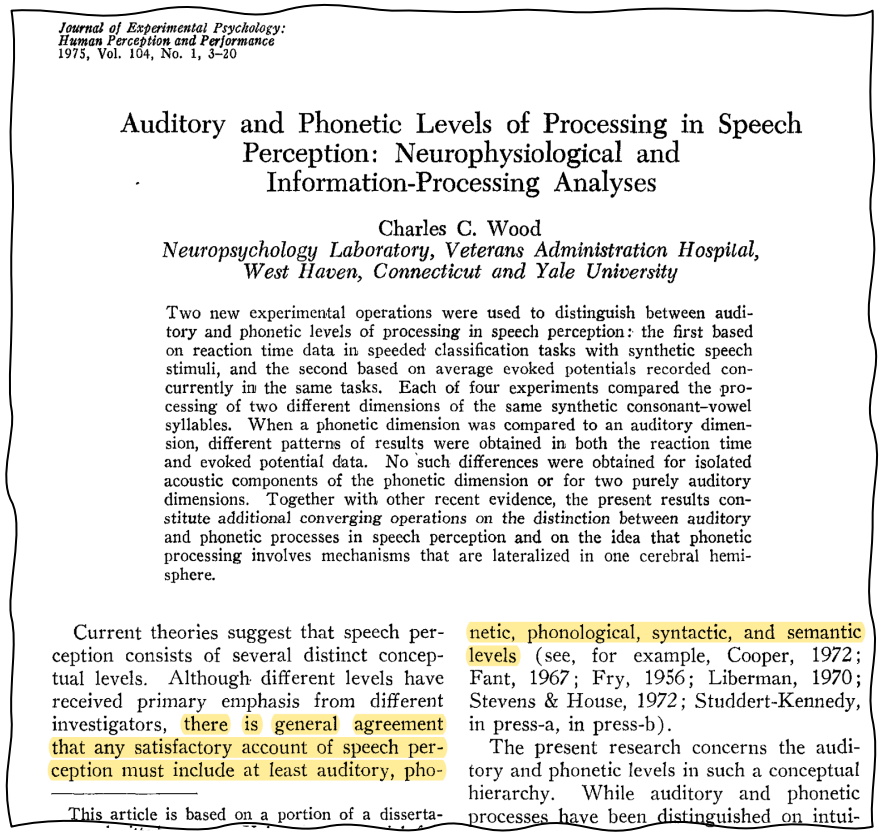
# Why multimodal?




**Engaging Content**
Engaging People

# What about humans?

- How do humans perceive speech?
- How do humans deal with noisy speech?
- Early 1970s focus on auditory processing
- Visual modality starts to emerge
  - gives place of articulation
  - helps in white noise



*Journal of Experimental Psychology: Human Perception and Performance*
1975, Vol. 104, No. 1, 3–20

Auditory and Phonetic Levels of Processing in Speech Perception: Neurophysiological and Information-Processing Analyses

Charles C. Wood
*Neuropsychology Laboratory, Veterans Administration Hospital, West Haven, Connecticut and Yale University*

Two new experimental operations were used to distinguish between auditory and phonetic levels of processing in speech perception: the first based on reaction time data in speeded classification tasks with synthetic speech stimuli, and the second based on average evoked potentials recorded concurrently in the same tasks. Each of four experiments compared the processing of two different dimensions of the same synthetic consonant–vowel syllables. When a phonetic dimension was compared to an auditory dimension, different patterns of results were obtained in both the reaction time and evoked potential data. No such differences were obtained for isolated acoustic components of the phonetic dimension or for two purely auditory dimensions. Together with other recent evidence, the present results constitute additional converging operations on the distinction between auditory and phonetic processes in speech perception and on the idea that phonetic processing involves mechanisms that are lateralized in one cerebral hemisphere.

Current theories suggest that speech perception consists of several distinct conceptual levels. Although different levels have received primary emphasis from different investigators, there is general agreement that any satisfactory account of speech perception must include at least auditory, phonetic, phonological, syntactic, and semantic levels (see, for example, Cooper, 1972; Fant, 1967; Fry, 1956; Liberman, 1970; Stevens & House, 1972; Studdert-Kennedy, in press-a, in press-b).

The present research concerns the auditory and phonetic levels in such a conceptual hierarchy. While auditory and phonetic processes have been distinguished on intui-

This article is based on a portion of a disserta-

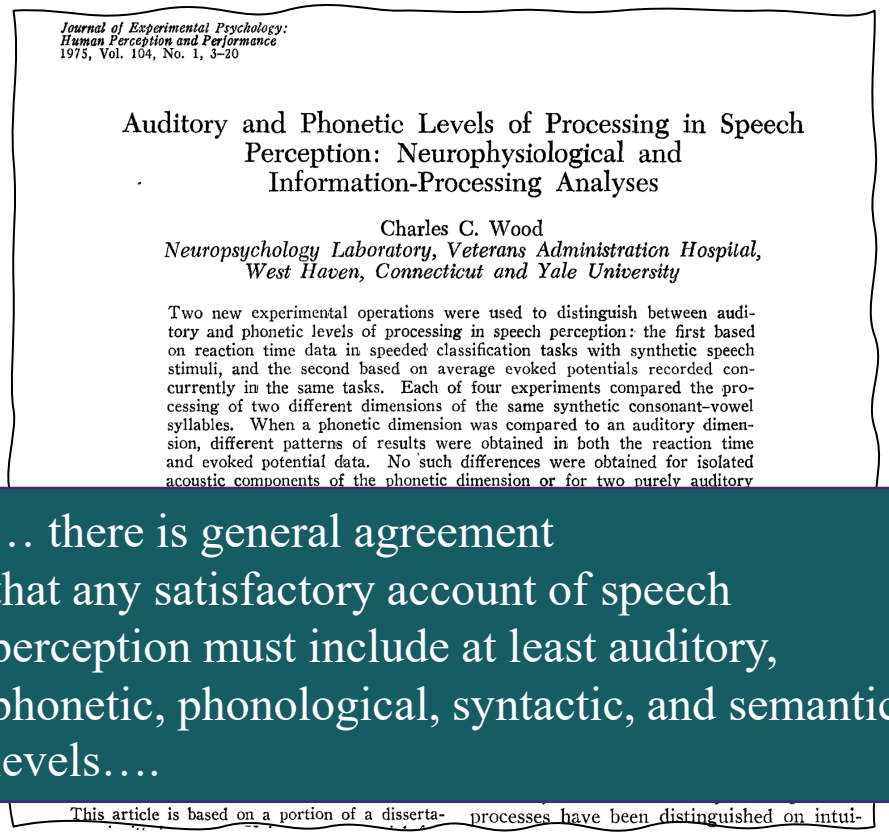
Wood, Charles C. "Auditory and phonetic levels of processing in speech perception: Neurophysiological and information-processing analyses." *Journal of Experimental Psychology: Human Perception and Performance* 1, no. 1 (1975): 3.

Binnie, C. A., Montgomery, A. A., and Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. Journal of Speech and Hearing Research, 17(4):619–630.

Dodd, B. (1977). The role of vision in the perception of speech. Perception, 6(1):31–40. PMID: 840618.

# What about humans?

- How do humans perceive speech?
- How do humans deal with noisy speech?
- Early 1970s focus on auditory processing
- Visual modality starts to emerge
  - gives place of articulation
  - helps in white noise



*Journal of Experimental Psychology: Human Perception and Performance*
1975, Vol. 104, No. 1, 3–20

Auditory and Phonetic Levels of Processing in Speech Perception: Neurophysiological and Information-Processing Analyses

Charles C. Wood
*Neuropsychology Laboratory, Veterans Administration Hospital, West Haven, Connecticut and Yale University*

Two new experimental operations were used to distinguish between auditory and phonetic levels of processing in speech perception: the first based on reaction time data in speeded classification tasks with synthetic speech stimuli, and the second based on average evoked potentials recorded concurrently in the same tasks. Each of four experiments compared the processing of two different dimensions of the same synthetic consonant–vowel syllables. When a phonetic dimension was compared to an auditory dimension, different patterns of results were obtained in both the reaction time and evoked potential data. No such differences were obtained for isolated acoustic components of the phonetic dimension or for two purely auditory

… there is general agreement that any satisfactory account of speech perception must include at least auditory, phonetic, phonological, syntactic, and semantic levels….

This article is based on a portion of a disserta- processes have been distinguished on intui-

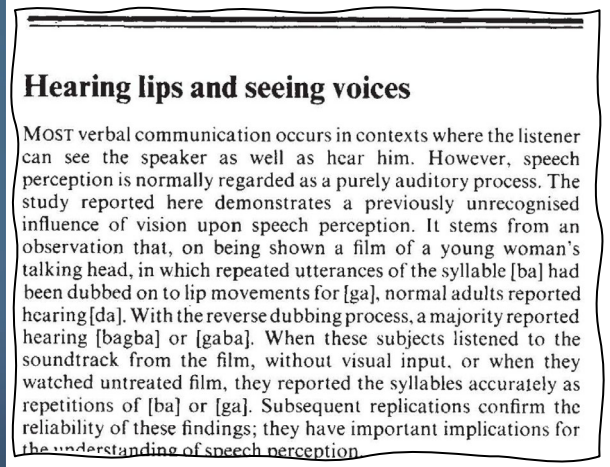
Wood, Charles C. "Auditory and phonetic levels of processing in speech perception: Neurophysiological and information-processing analyses." *Journal of Experimental Psychology: Human Perception and Performance* 1, no. 1 (1975): 3.

Binnie, C. A., Montgomery, A. A., and Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. Journal of Speech and Hearing Research, 17(4):619–630.

Dodd, B. (1977). The role of vision in the perception of speech. Perception, 6(1):31–40. PMID: 840618.

# The McGurk Effect


McGurk, Harry, and John MacDonald. "Hearing lips and seeing voices." *Nature* 264.5588 (1976): 746-748.



**Hearing lips and seeing voices**

MOST verbal communication occurs in contexts where the listener can see the speaker as well as hear him. However, speech perception is normally regarded as a purely auditory process. The study reported here demonstrates a previously unrecognised influence of vision upon speech perception. It stems from an observation that, on being shown a film of a young woman's talking head, in which repeated utterances of the syllable [ba] had been dubbed on to lip movements for [ga], normal adults reported hearing [da]. With the reverse dubbing process, a majority reported hearing [bagba] or [gaba]. When these subjects listened to the soundtrack from the film, without visual input, or when they watched untreated film, they reported the syllables accurately as repetitions of [ba] or [ga]. Subsequent replications confirm the reliability of these findings; they have important implications for the understanding of speech perception.

Audio “ba” + Video “fa”
Perceive “fa”

Audio “ba” + Video “ba”
Perceive “ba”

Full video on: https://www.youtube.com/watch?v=2k8fHR9jKVM&t=62s

# How can the visual side help?

- Maximum benefit to speech intelligibility
  - auditory SNR range 6dB to -5dB
- Localisation
- Voice activity
- Asynchrony
- Distant speech

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In Dodd, B. and Campbell, R., editors, Hearing by eye: The psychology of lip-reading., pages 3–51. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US.

Tuomainen, Jyrki, Tobias S. Andersen, Kaisa Tiippana, and Mikko Sams. "Audio–visual speech perception is special." *Cognition* 96, no. 1 (2005): B13-B22.

# A wider concept of multimodality

- Speech recognition is a narrow viewpoint
- Human conversation
  - The words we use…
  - Or how we say then…
  - Hand gestures, head nods, eye gaze
  - Back-channels
  - Open mouth
- Effortlessness in combining cues

Source: https://neurosciencenews.com/gestures-visual-linguistics-12063/

# Speech is multimodal

- Non-verbal aspects of interaction
- Crucial to take multimodal approach for robustness

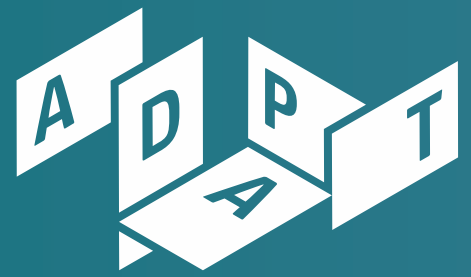
- Asynchrony in speech

- Cues change at different rates

# AV-Align

Teaching DNNs to learn from the visual side of speech too

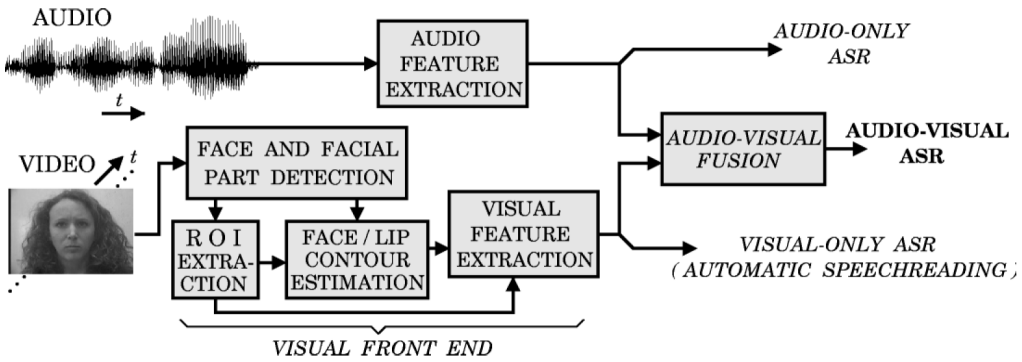
(Acknowledging PhD work of George Sterpu)




**Engaging Content**
Engaging People

# AVSR – early work

- Good visual representation
  - Shape, appearance
  - PCA, DCT, HAAR, AAM
- Integration of audio and visual signals
  - Early, intermediate, late
  - Feature fusion to decision fusion
- Tasks
  - Isolated words, connected digits
  - Lack of datasets




Hennecke, M. E., Stork, D. G., and Prasad, K. V. (1996). Visionary speech: Looking ahead to practical speechreading systems. In Stork, D. G. and Hennecke, M. E., editors, Speechreading by Humans and Machines: Models, Systems, and Applications, pages 331–349. Springer Berlin Heidelberg, Berlin, Heidelberg.

Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, *91*(9), 1306-1326.

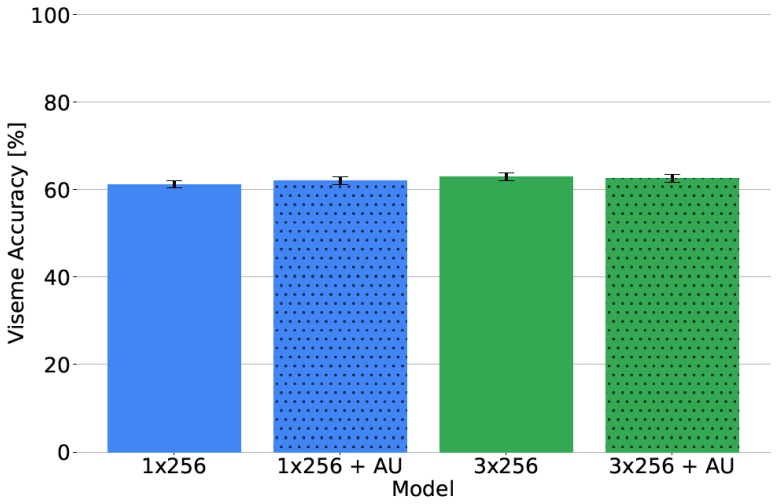
# LRS2

- Data from BBC
  - 96318 examples for pre-training
  - 45839 for training
  - 1243 for testing
- More challenging
  - Head pose, illumination changes, low image resolution, 15000 words
- Note LRS3 is less challenging

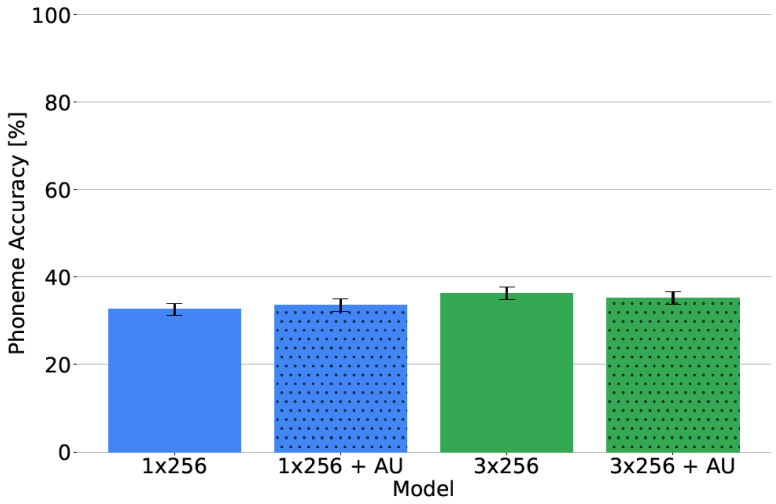
BBC and University of Oxford (2017). The Oxford-BBC Lip Reading Sentences 2 (LRS2) Dataset. http://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html.

# Lipreading on LRS2

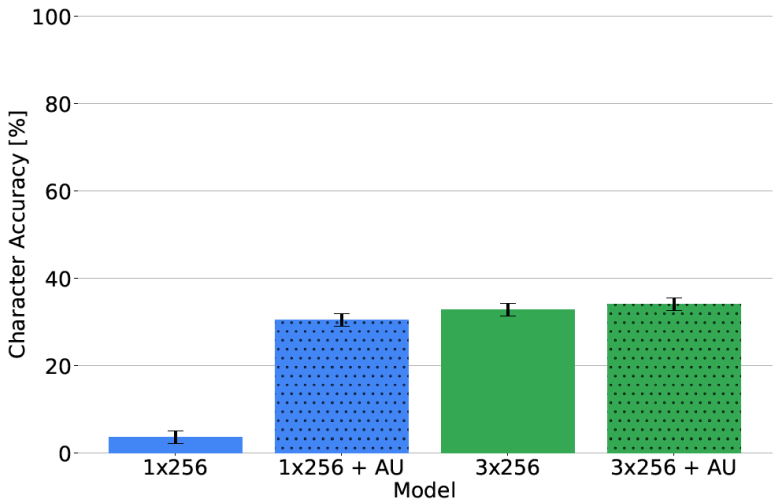
- LSTMs, what unit?




**(a)** *Lipreading Visemes*



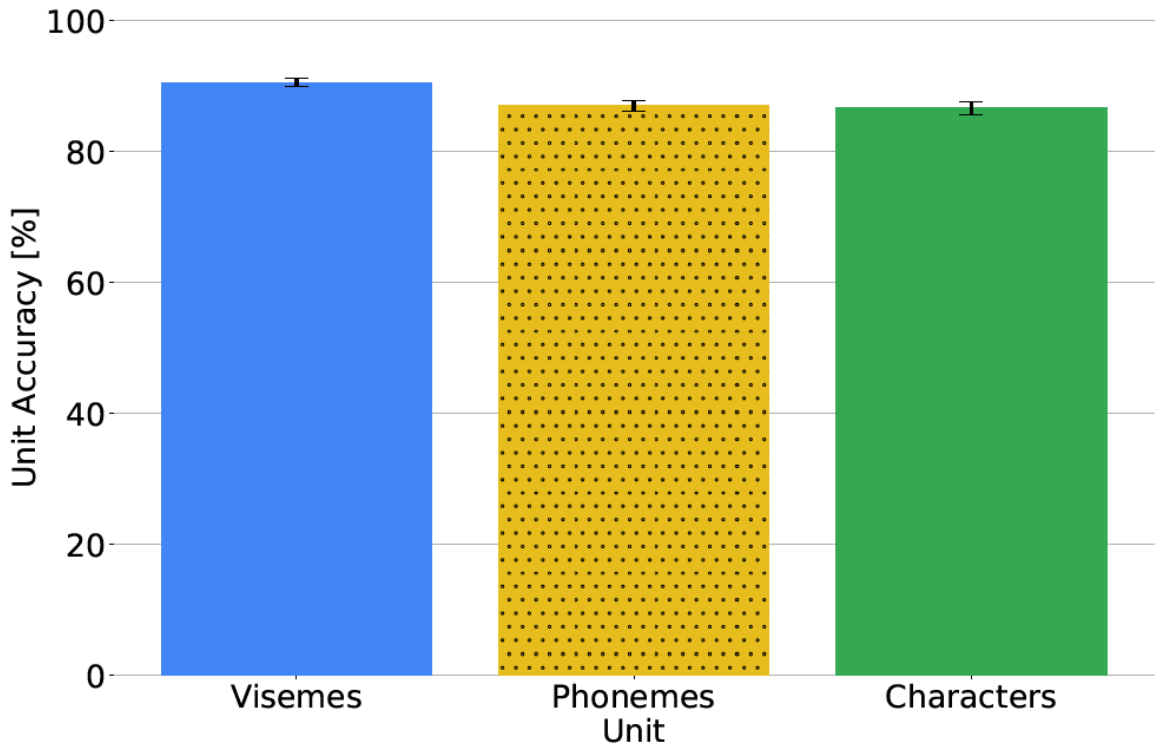

**(b)** *Lipreading Phonemes*




**(c)** *Lipreading Characters*

# Lipreading on LRS2

- Audio-only LSTM. Too good!!




**(d)** *Audio-only performance on clean speech*

# Motivation

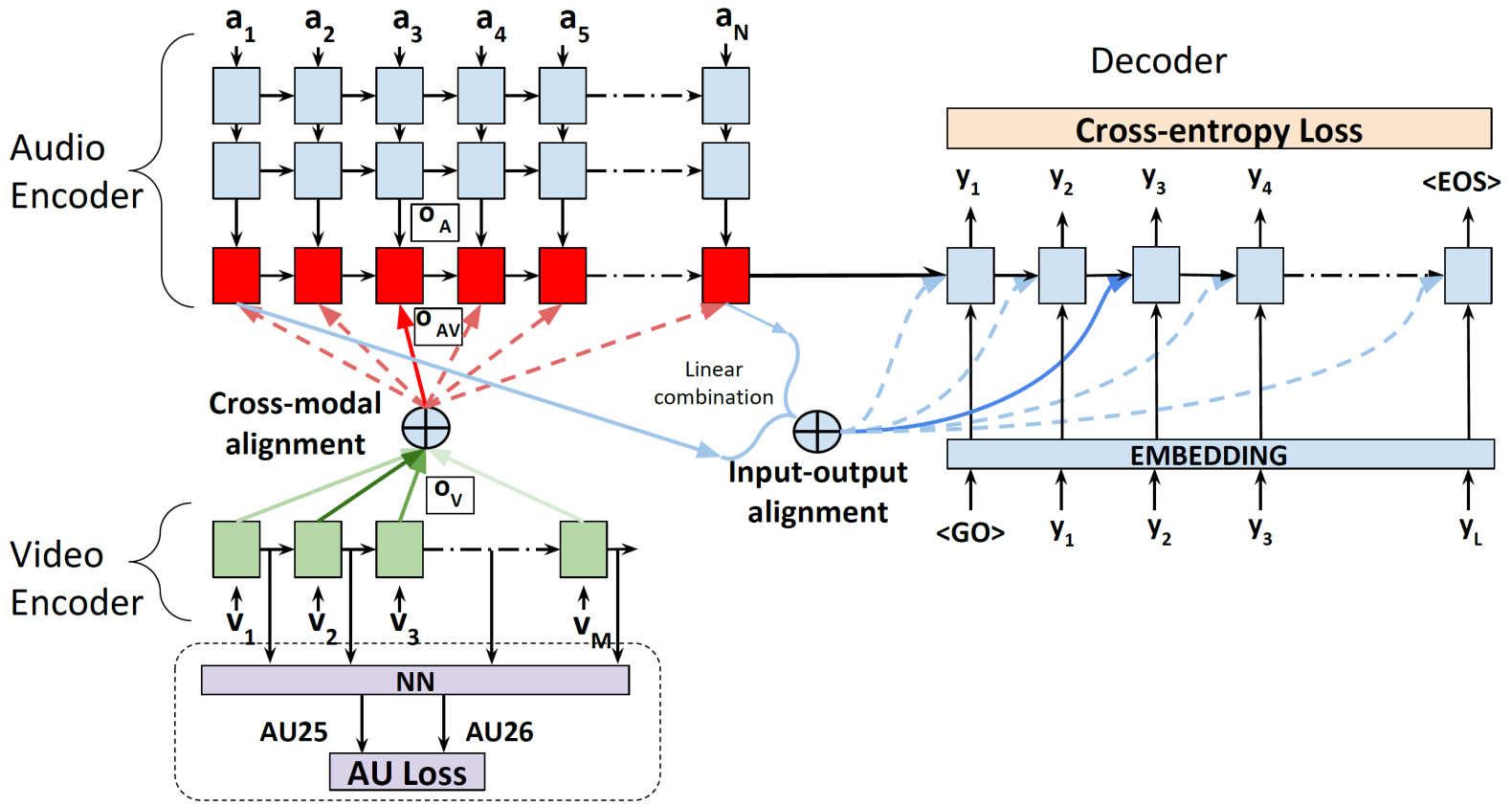
- Hypothesis?
- *The visual modality plays a complimentary role to the audio in speech and we need to teach that to a neural net*

# To AV-Align

- Maybe visemes are not so useful?
- Visual modality limited linguistically
- DNN-based multimodal systems
  - encode each modality separately
  - representations fused at decoding (Afouras et al., 2018; Chung et al., 2017; Petridis et al., 2018).
- *… the acoustic representations of speech are altered by the visual representations during a multimodal encoding process, before decoding starts. What the system sees, influences what it hears…*

# AV-Align Strategy

- The top layer cells of the Audio Encoder take audio representations from a stack of LSTM layers ($o_A$) as inputs and attend to the top layer outputs of the Video Encoder ($o_V$ , only one layer shown), producing the crossmodal alignment.
- The Decoder receives the fused Audio-Visual representations ($o_{AV}$ ), producing an input-output alignment through a second attention mechanism.
- Dashed lines depict inactive states in a hard selection process, whereas shaded lines stand for a soft selection mechanism.




Sterpu, George, Christian Saam, and Naomi Harte. "How to teach DNNs to pay attention to the visual modality in speech recognition." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020): 1052-1064.

# Compare to WLAS (Watch Listen Attend and Spell)

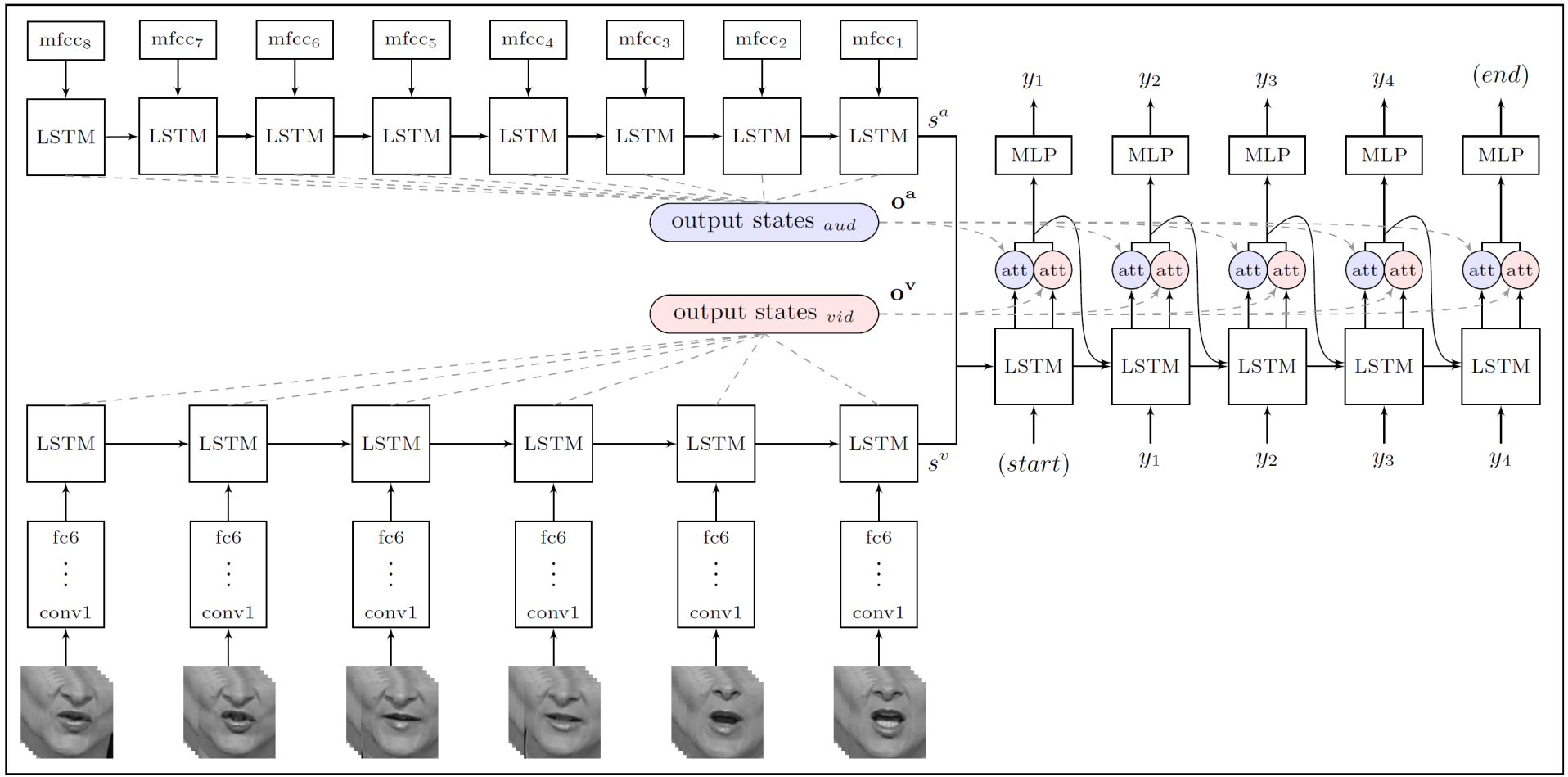


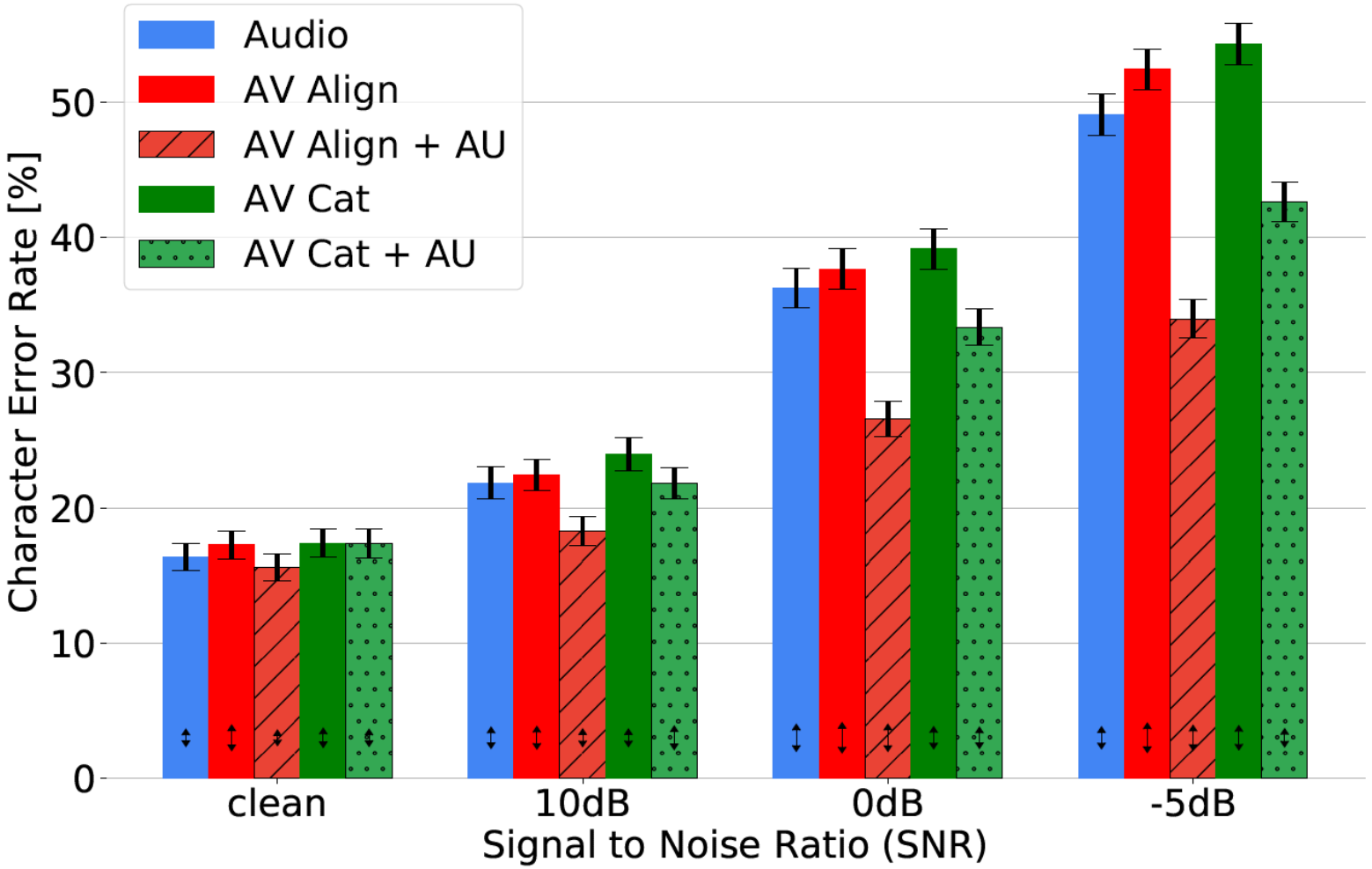

Figure 1. *Watch, Listen, Attend and Spell* architecture. At each time step, the decoder outputs a character $y_i$, as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2017). Lip reading sentences in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

# Evaluation set-up

- Audio
  - 16,000 Hz
  - log magnitude spectrogram, frame length 25ms, 10ms stride
  - 30 mel bins 80Hz to 11,025Hz
  - Cafeteria noise
- Video
  - OpenFace to detect and align faces (discard ~ 2.77%)
  - RGB images of the lip regions downsampled to 36x36 pixels
  - ResNet CNN visual feature vector of 128 units per frame
- Training in 4 stages
  - clean speech, 10db, 0db, -5db
- https://github.com/georgesterpu/avsr-tf1

# AV-Align performance



# Learn to exploit the asynchrony?

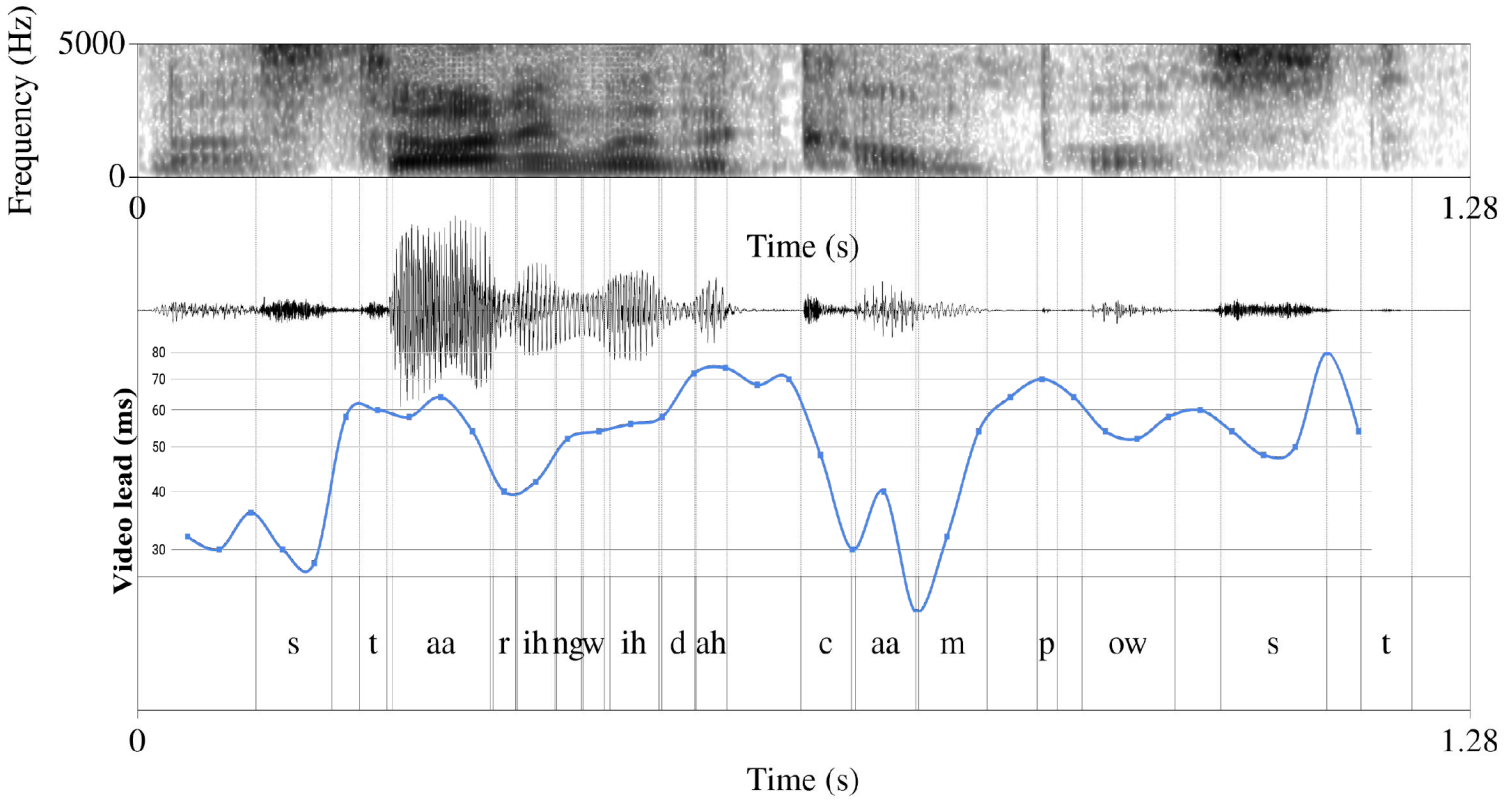
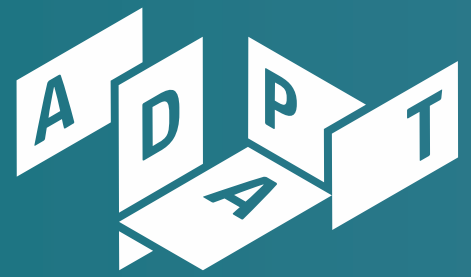



**Figure 4.21:** *Phonetic analysis of the modality lags predicted by AV Align for the sentence* "Starting with the compost"*, showing the speech spectrogram, waveform, modality lag, and transcription. The delay between modalities is estimated by fitting a normal distribution for each column (audio frame) of the cross-modal alignment matrix and selecting the mean.*

# Neural Turn-Taking prediction

Combining modalities to predict turns

Acknowledging PhD work of Matt Roddy




**Engaging Content**
Engaging People

# Turn taking – MapTask example

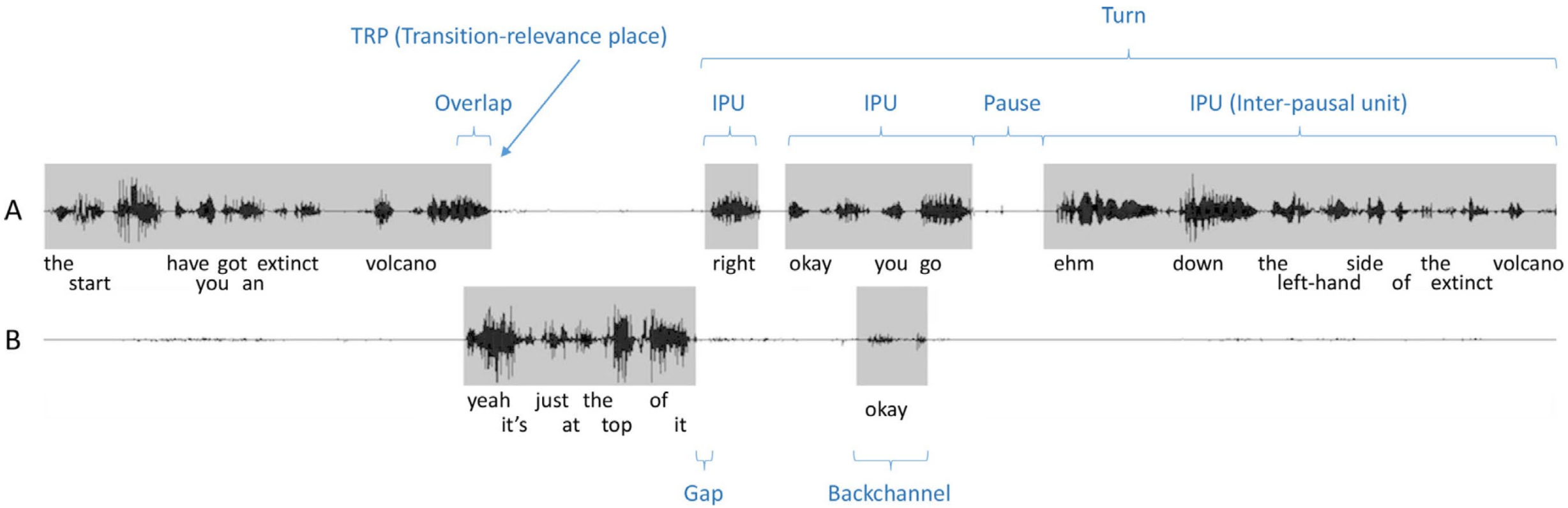



Diagram from: Skantze, Gabriel. "Turn-taking in conversational systems and human-robot interaction: a review." *Computer Speech & Language* (2020): 101178.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., ... & Weinert, R. (1991). The HCRC map task corpus. *Language and speech*, *34*(4), 351-366.

# Gaps in real conversations are short

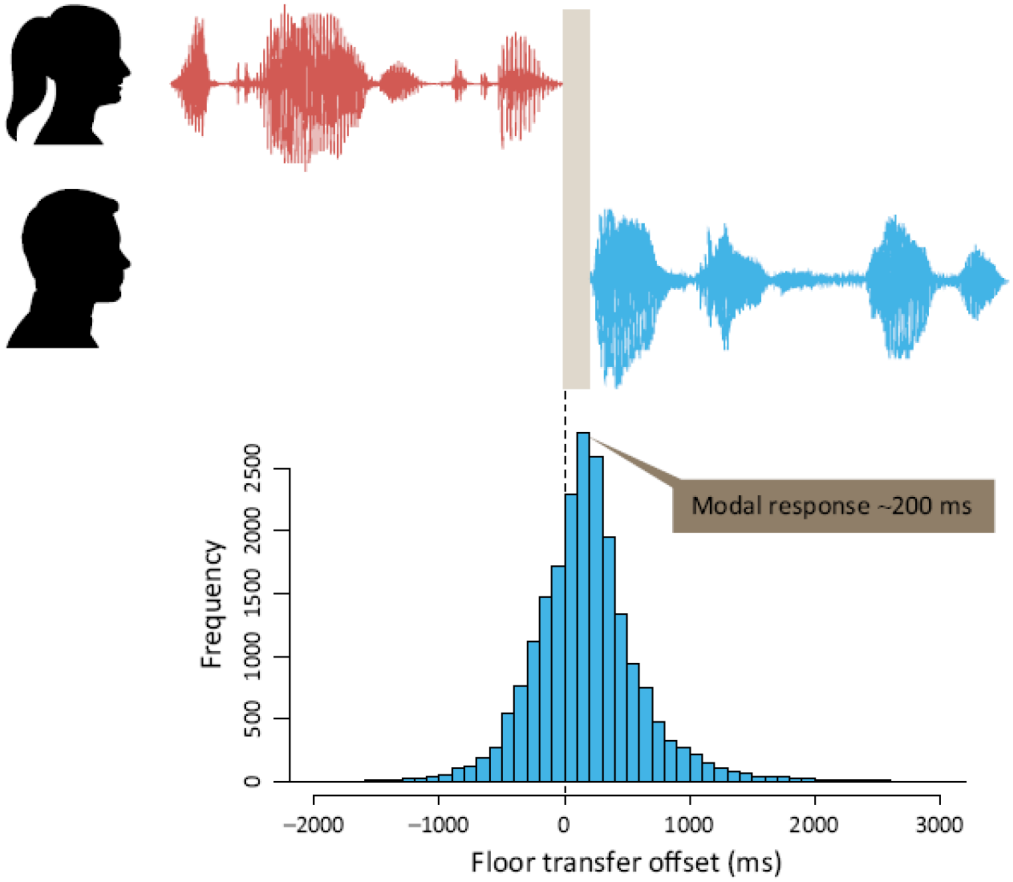



Figure 1.1: Histogram of floor transfer offset timings. From (Levinson, 2016)

# Dialogue systems

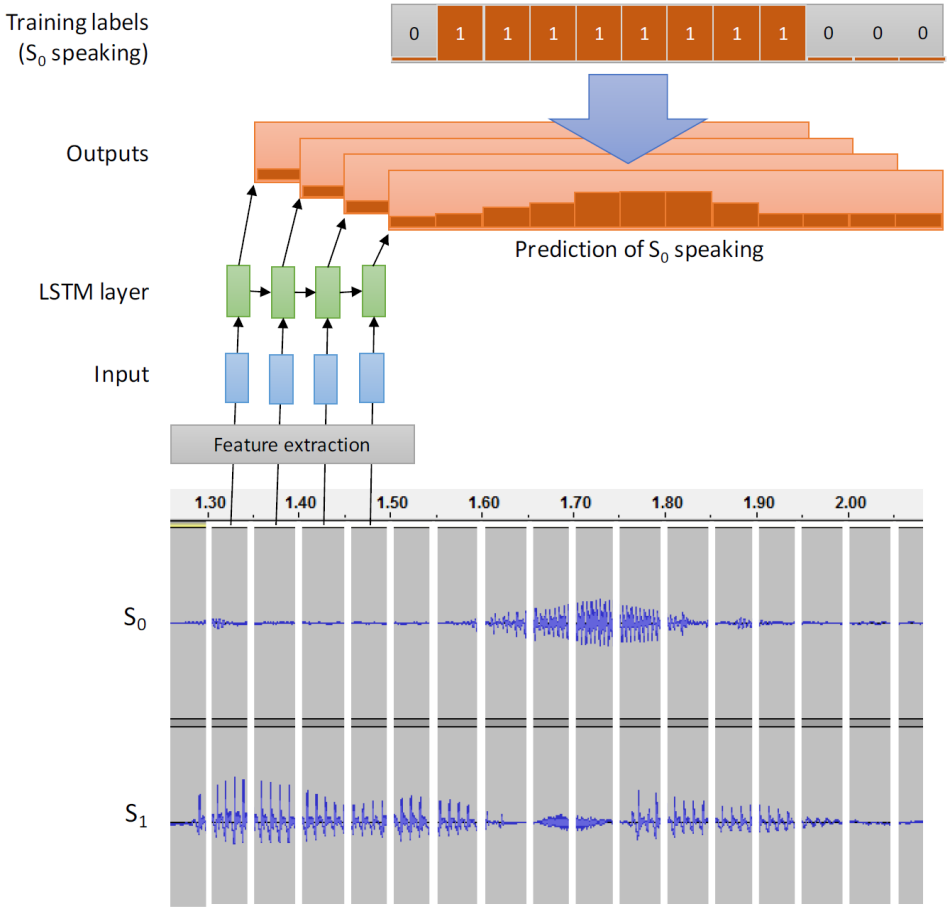
- Detecting silences and equating to end of turn is not widely useful
- Humans interpret many cues
  - Multiple modalities
  - In parallel
  - Additive effect
- Turn yielding
- Turn holding
- Turn initial

**Table 1** (Source Skantze 2021)
Typical turn-final cues found in studies of English conversation.

| | Turn-yielding cues | Turn-holding cues |
|---|---|---|
| Verbal | Syntactically complete | Syntactically incomplete, Filled pause |
| Prosody | Rising or falling pitch, Lower intensity | Flat pitch, Higher intensity |
| Breathing | Breathe out | Breathe in |
| Gaze | Looking at addressee | Looking away |
| Gesture | Terminated | Non-terminated |

# Continuous turn-taking model

- LSTM
- Acoustic and POS features
- 50 ms frame
- probabilistic predictions for individual future frames within a window of length N
- Continuous predictions rather than detecting turn switch based on prior events





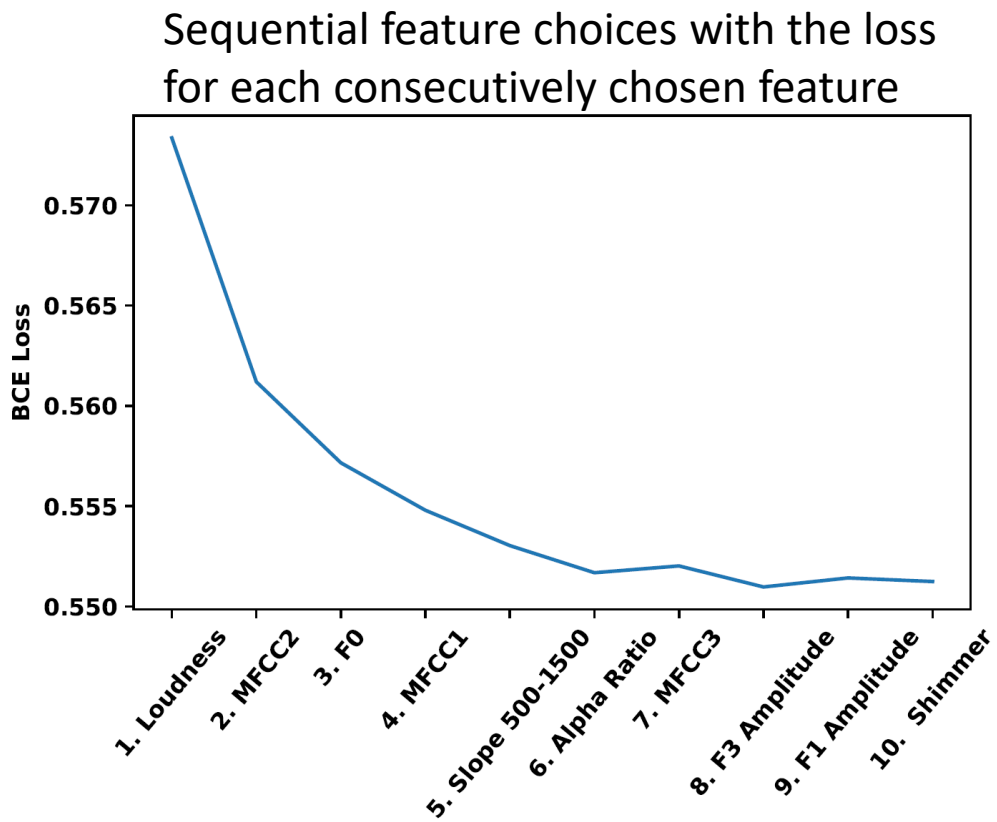
Skantze, G. (2017, August). Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue* (pp. 220-230).

# More features

- Extended Skantze 2017 model with some modifications
  - Removed truncated back-propagation through time,  used binary cross-entropy (BCE) loss
- Acoustic Features
  - eGeMAPs feature set
- Linguistic Features
  - Part-of-Speech (POS), and word embeddings
- Phonetic Features
  - bottleneck layer output of  DNN trained to classify senones (tied tri-phone states)
- Voice Activity
- HCRC map task corpus

| | |
|---|---|
| Frequency | pitch; jitter; centre frequencies of formants 1, 2, and 3; bandwidth of first formant |
| Energy | loudness; shimmer; harmonics-to-noise ratio (HNR) |
| Spectral | MFCCs 1-4; spectral flux; alpha ratio; Hammarberg Index; spectral slope 0-500 Hz and 500-1500 Hz; relative energy of formants 1, 2, and 3; |

# Insights



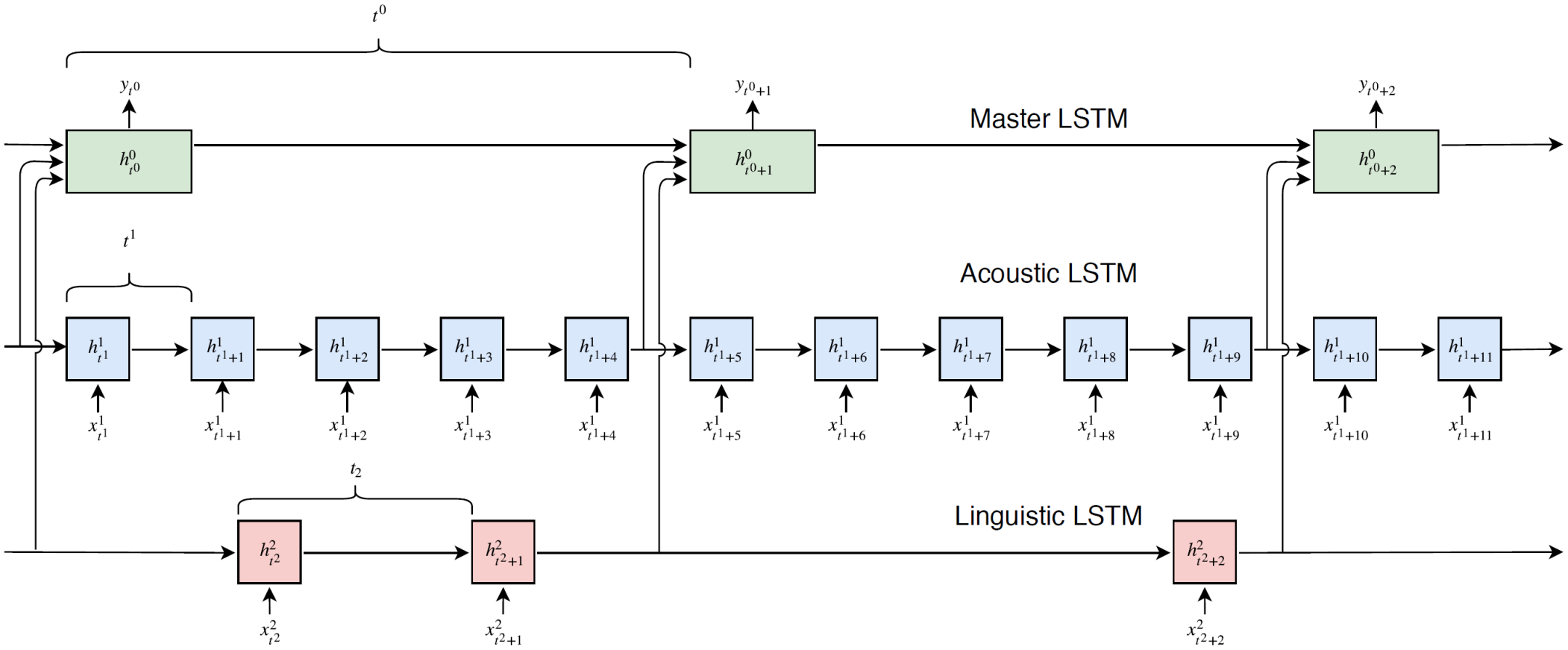

- Lexical embeddings always outperformed syntactic features
- Acoustic and lexical features
- But generally linguistic features are better in the literature?
- How is modality-specific information temporally represented?

Roddy, M., Skantze, G., Harte, N. (2018) Investigating Speech Features for Continuous Turn-Taking Prediction Using LSTMs. Proc. Interspeech 2018, 586-590

- Hierarchical structure of language during a speaking turn
  - phonemes to create words
  - words to create IPUs
  - IPUs to create turns


- Tricky for continuous turn taking models
- 50ms frame?
- Multiscale RNN architecture
  - modalities modelled in separate sub-network LSTMs
  - independent timescales
- HCRC map-task corpus - linguistic + acoustic
- Mahnob Mimicry Database - visual + acoustic

# Multiscale RNN architecture




Roddy, Matthew, Gabriel Skantze, and Naomi Harte. "Multimodal continuous turn-taking prediction using multiscale RNNs." In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pp. 186-190. 2018.

www.github.com/mattroddy/lstm_turn_taking_prediction

# Features

- Acoustic
  - As before but 10 and 50 ms
- Linguistic
  - 64-length word embeddings as before
  - 10 ms, 50ms, asynchronous
- Visual
  - gaze direction predictions (x, y, z) for each eye and confidence score from OpenFace

# MapTask – audio + linguistic

- Metric F1 score on prediction Hold or Shift at pauses
- Onset – short or long
- Multiscale beneficial
- Slower rate linguistic better
- Combined at different rates better

Predictions at 50 and 500ms pauses

Predictions at onset

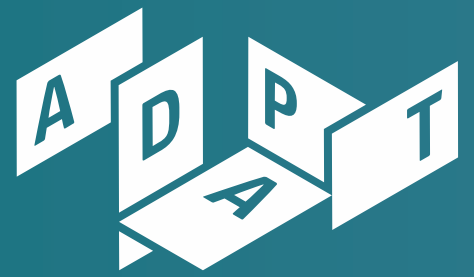
| | BCE loss | f1 50ms | f1 500ms | f1 onset |
|---|---|---|---|---|
| No Subnets (Early Fusion) | | | | |
| (1) Acous 50ms | 0.5456 | 0.7907 | 0.8165 | 0.7926 |
| (2) Acous 10ms | 0.5351 | 0.8154 | 0.8428 | 0.8126 |
| (3) Ling 50ms | 0.5779 | 0.7234 | 0.7547 | 0.7249 |
| (4) Ling Asynch | 0.5839 | 0.7101 | 0.7341 | 0.7174 |
| (5) Ling 10ms | 0.5823 | 0.7072 | 0.7391 | 0.7111 |
| (6) Acous 50ms Ling 50ms | 0.5411 | 0.7957 | 0.8354 | 0.8101 |
| (7) Acous 10ms Ling 10ms | **0.5321** | **0.8194** | **0.8465** | **0.8141** |
| One Subnet | | | | |
| (8) Acous 50ms Ling 50ms | 0.5414 | 0.7922 | 0.8366 | 0.8020 |
| (9) Acous 10ms Ling 10ms | **0.5317** | **0.8237** | **0.8480** | **0.8128** |
| Two Subnets (Multiscale) | | | | |
| (10) Acous 50ms Ling 50ms | 0.5420 | 0.7916 | 0.8303 | 0.8019 |
| (11) Acous 10ms Ling 50ms | ***0.5291*** | ***0.8323*** | 0.8526 | ***0.8236*** |
| (12) Acous 50ms Ling Asynch | 0.5416 | 0.7949 | 0.8385 | 0.7993 |
| (13) Acous 10ms Ling Asynch | 0.5296 | 0.8307 | ***0.8553*** | 0.8232 |
| (14) Acous 10ms Ling 10ms | 0.5310 | 0.8285 | 0.8470 | 0.8189 |

# Mahnob – audio + visual

- Faster 58Hz gaze features good potential
- Visual features matter more in different types of switches

| | BCE loss | f1 50ms | f1 500ms | f1 onset |
|---|---|---|---|---|
| No Subnets (Early Fusion) | | | | |
| (1) Acous 50ms | 0.4433 | 0.8665 | 0.9230 | 0.8668 |
| (2) Acous 10ms | **0.4348** | ***0.8851*** | **0.9343** | **0.8685** |
| (3) Visual 50ms | 0.5840 | 0.7858 | 0.8154 | 0.6445 |
| (4) Visual 58Hz | 0.5941 | 0.7726 | 0.8031 | 0.6560 |
| (5) Acous 50ms Visual 50ms | 0.4497 | 0.8651 | 0.9159 | 0.8526 |
| Two Subnets (Multiscale) | | | | |
| (6) Acous 50ms Visual 50ms | 0.4443 | 0.8637 | 0.9198 | 0.8711 |
| (7) Acous 10ms Visual 50ms | 0.4337 | **0.8840** | ***0.9347*** | ***0.8784*** |
| (8) Acous 50ms Visual 58Hz | 0.4437 | 0.8634 | 0.9216 | 0.8721 |
| (9) Acous 10ms Visual 58Hz | ***0.4332*** | 0.8831 | 0.9343 | 0.8762 |

# AI for Multimodal Perception?

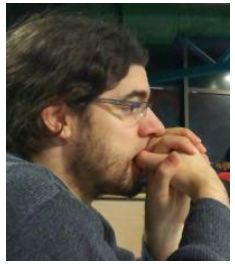



**Engaging Content**
Engaging People

# Take-home…

- Speech is multimodal
- Consider how the modalities interplay
- Yields most suitable neural architectures
- Data won’t always save you
- Less is more!

# Thank you!



Dr. Sebastien Le Maguer

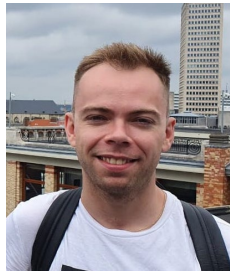
Dr. Justine Reverdy

Mark Anderson

Ayushi Pandey

Sam Kotey

Ed Storey



Sam Russell

Dr. George Sterpu
(Xperi)

Dr. Matt Roddy
(Cognito Corp)

nharte@tcd.ie

Engaging Content
Engaging People

www.adaptcentre.ie

FUNDED BY:
Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
European Union
European Regional Development Fund
Science Foundation Ireland
sfi For what's next