Visual speech processing: aligning terminologies for better understanding

Helen L Bear and Sarah Taylor
helen@uel.ac.uk s.l.taylor@uea.ac.uk

Speechreading or lipreading?

Despite commonly being used interchangeably the terms speechreading and lipreading have subtle but distinctive definitions.

Figure 1: Regions of an image used in speech reading.

Speechreading is what human lipreaders do. It uses information from the whole face and body since knowledge of facial expression, gaze, and body gestures often helps to provides semantic context that makes decoding speech easier.

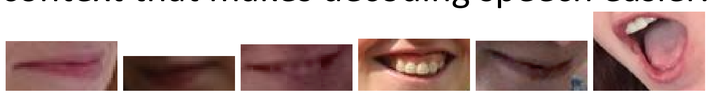
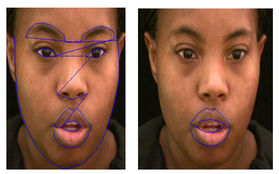


Figure 2: Regions of an image used in lipreading.

Lipreading is the interpretation of speech from the motion of the lips alone.

Tracking or classifying?



It is hard to track the lips in isolation as the lips have no skeletal structure and a deformable surface.

Figure 4: Examples of shapes for a full face tracking (left), and a lip only track (right).

Track	Features	System
Full face	Full face	Speechreading
Full face	Lips only	Lipreading
Lips only	Lips only	Lipreading

Speaker independence

Speaker independence in machine lipreading is achieved when classification models recognise talkers not contained within the training set.

E.g, dataset A contains 1000 utterances by speaker X. To build a speaker dependent lipreading system we use 800 training and 200 test samples. To achieve speaker independence, it is not sufficient to only separate specific utterances of a speaker, i.e. sentences 1 to n for speaker X to train, and sentences n+1 to 1000 for test. Alternatively, if we have speaker X and speaker Y in a dataset, each with 1000 utterances, we can use the 800 speaker X training samples train, and we test on 200 speaker Y samples & vice versa. This is speaker independent lipreading, Figure 5.

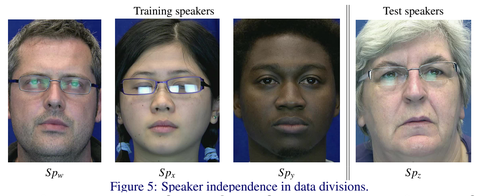


Figure 5: Speaker independence in data divisions.

For classification methods that require the data to be split into train, validation, and test sets, the validation set can contain speakers from the training set and new speakers, but speakers must remain distinct from the test set if speaker independence is the goal, Figure 6.

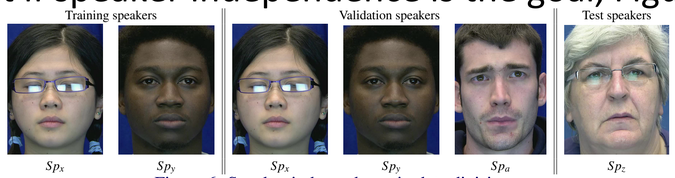


Figure 6: Speaker independence in data divisions.

Note that for any duplicate speakers in both training and validation sets, one must split samples, i.e. sample 1 for speaker X can only be a either training or validation sample.

Scoring metrics

Methods of reporting on the performance of machine lipreading have been adopted from audio speech recognition systems.

$$C = \frac{N-D-S}{N} \quad or \quad C\% = \frac{N-H}{N} \times 100\%$$

$$A = \frac{N-D-S-I}{N} \quad or \quad A\% = \frac{N-H-I}{N} \times 100\%$$

$$ER\% = \frac{D-S-I}{N} \times 100\%$$

N= total number of labels in the ground truth, D= the number of deletion errors, S= the number of substitution errors, I=the number of insertion errors. H=D+S.

But these also vary by system recognition units: words, or phonemes, or visemes. This unit selection is chosen for each classifier and language net phases of a lipreading system. Thus we suggest the notation;

$$M_{cu} \quad || \quad M^{nu}$$

M is the metric {*A*,*C*, *ER*}, subscript denotes a classifier score, or a superscript denotes a network score

Top 5 or Top 1 match.

Interclass variation in a set of phonemes or visemes is much smaller than, for e.g. that within a set of categorized images. This means a classified output can easily have a significantly different meaning which confuses conversation talkers. So we recommend only top one scores are reported.

UEL University of East London

University of East Anglia