Lipreading in the Wild

Joon Son Chung and Andrew Zisserman

https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16/chung16.pdf

TABLE OF CONTENTS

1. EXISTING LIPREADING DATASETS

Name	Env.	Output	I/C	# class	# subj.	Best perf.
AVICAR [19]	In-car	Digits	C	10	100	37.9% [7]
AVLetter [22]	Lab	Alphabet	I	26	10	43.5% [43]
CUAVE [27]	Lab	Digits	I	10	36	83.0% [26]
GRID [4]	Lab	Words	C	8.5*	34	79.6% [39]
OuluVS1 [43]	Lab	Phrases	I	10	20	89.7% [28]
OuluVS2 [1]	Lab	Phrases	I	10	52	73.5% [44]
OuluVS2 [1]	Lab	Digits	C	10	52	-
BBC TV	TV	Words	C	333/500	1000+	-

Table 1. Existing lip reading datasets. **I** for **I**solated (one word, letter or digit per recording); **C** for **C**ontinuous recording. The reported performance is on speaker-independent experiments. (* For GRID [4], there are 51 classes in total, but the first word in a phrase is restricted to 4, the second word 4, etc. 8.5 is the average number of possible classes at each position in the phrase.)

2. BUILDING THE DATASET

PROBLEM: Extract 1000+ video clips of 500+ words

Sub-problems: Subtitle-audio timing alignment, Active speaker identification/tracking

SOLUTION:

(i) obtain a temporal alignment of the spoken audio with a text transcription (broadcast as subtitles with the program)

(ii) obtain a spatio-temporal alignment of the lower face for the frames corresponding to the word sequence

(iii) determine that the face is speaking the words (i.e. that the words are not being spoken by another person in the shot)

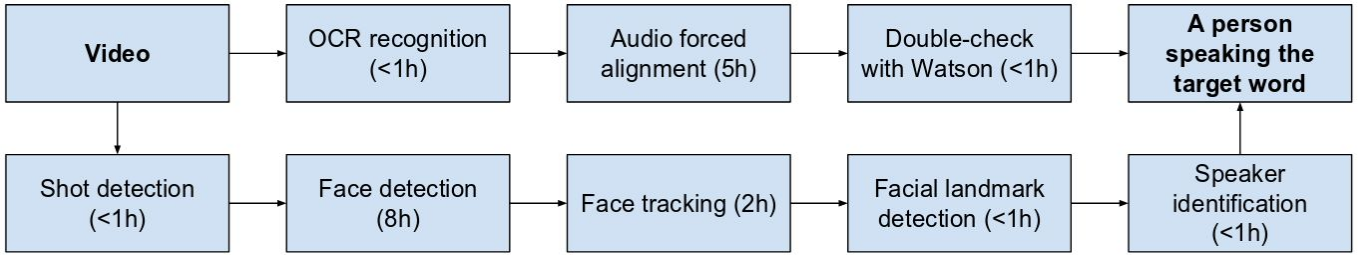

Fig. 2. Pipeline to generate the text and visually aligned dataset. Timings are for a one-hour video.

2. BUILDING THE DATASET

1. **Stage 1. Selecting program types.**
2. **Stage 2. Subtitle processing and alignment**

- The Penn Phonetics Lab Forced Aligner [9, 41] (based on the open-source HTK toolbox [40]) is used to force-align the subtitle to the audio signal
- The noisy labels are filtered by double-checking against the commercial IBM Watson Speech to Text service

3. **Stage 3. Shot boundary detection, face detection, and tracking**

dlib, KLT tracker

4. **Stage 4. Facial landmark detection and speaker identification**

- The ‘openness’ of the mouth is measured on every frame using the distance between the top and the bottom lip, normalised w.r.t. the size of the face -> LinearSVM

5. **Stage 5. Compiling the training and test data.**

2. BUILDING THE DATASET

Fig. 5. One-second clips that contain the word '*about*'. Top: male speaker, bottom: female speaker.

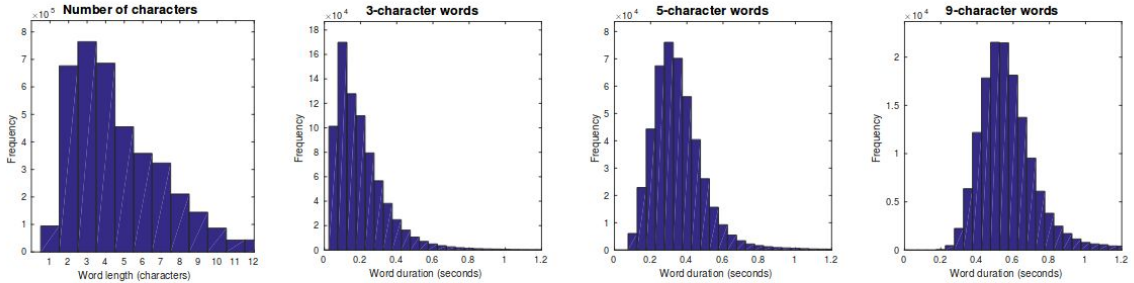

Fig. 6. Word statistics. Regardless of the actual duration of the word, we take a 1-second clip for training and test.

3. NETWORK ARCHITECTURE AND TRAINING

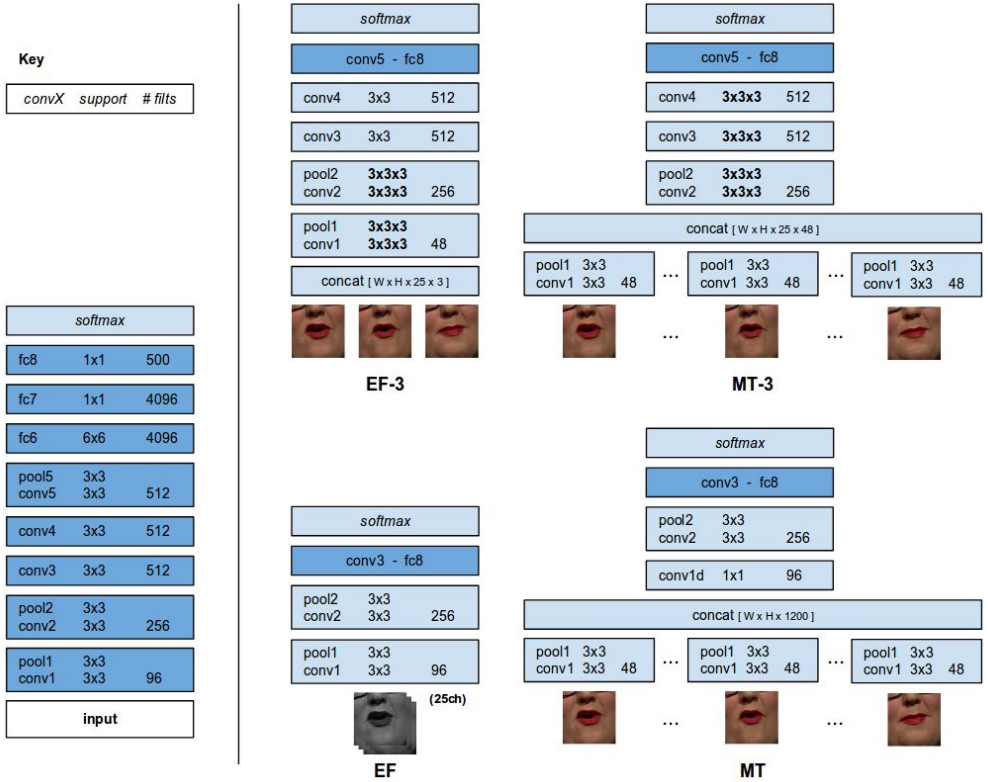

Fig. 7. CNN architectures. **Left:** VGG-M architecture that is used as a base. **Right:** **EF-3:** 3D Convolution with Early Fusion; **MT-3:** 3D Convolution with Multiple Towers; **EF:** Early fusion; **MT:** Multiple Towers.

Also,
DATA AUGMENTATION

4. EXPERIMENTS

Results. As discussed in Section 3.1, the **MT-3** and **MT** variants have the advantage of being more tolerant to registration errors compared to their early fusion counterparts. The results in Table 4 and Figure 8 confirm this, where we see a modest (3.2% on average for *top-1*) but consistent improvement in performance across the experiments. The performance of 3D ConvNets fall short of the 2D architectures by an average of around 14%.

The recall curves in Figure 8 rise sharply for all models at low-K; the *top-10* figure for the **EF** and **MT** models being over 85%, despite the modest *top-1* figure of around 60%. This is a result of ambiguities in lip reading, which we will discuss next.

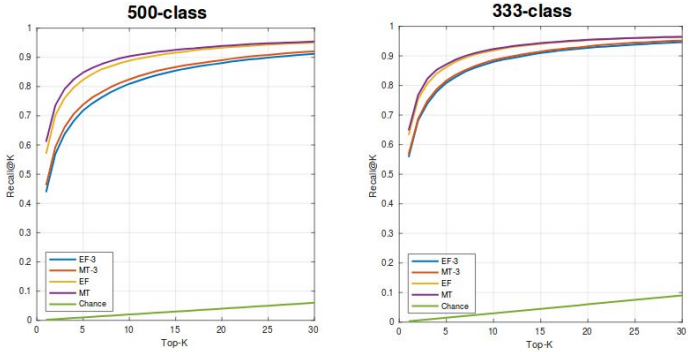

Fig. 8. Recall vs Rank curves for the word classification.

Net	500-class			333-class	
	Top-1	Top-10	ED	Top-1	Top-10
EF-3	43.9%	81.0%	3.13	55.7%	87.9%
MT-3	46.2%	82.4%	2.97	56.8%	88.7%
EF	57.0%	88.8%	2.32	63.2%	91.8%
MT	**61.1%**	**90.4%**	**2.06**	**65.4%**	**92.3%**

	OuluVS1	OuluVS2
	Top-1	Top-1
[29]	81.8%	-
[44]	85.6%	73.5%
[28]	89.7%	-
MT	**91.4%**	**93.2%**

Table 4. Word classification results. **Left:** On the BBC data for the four different architectures. **ED** is the edit distance. **Right:** On OuluVS1 and OuluVS2 (short phrases, frontal view).

500-class			333-class		
0.32	BENEFITS	BENEFIT	0.30	BORDER	IMPORTANT
0.31	QUESTIONS	QUESTION	0.29	PROBABLY	PROBLEM
0.31	REPORT	REPORTS	0.27	TAKING	TAKEN
0.31	BORDER	IMPORTANT	0.25	PERSONAL	PERSON
0.31	AMERICA	AMERICAN	0.23	CLAIMS	GAMES
0.29	GROUND	AROUND	0.22	AROUND	GROUND
0.28	RUSSIAN	RUSSIA	0.21	TONIGHT	NIGHT
0.28	FIGHT	FIGHTING	0.21	PROBLEM	PROBABLY
0.26	FAMILY	FAMILIES	0.19	SEVERAL	SEVEN
0.26	AMERICAN	AMERICA	0.19	CHALLENGE	CHANGE
0.26	BENEFIT	BENEFITS	0.18	PRICES	PERSON
0.25	ELECTIONS	ELECTION	0.18	WARNING	MORNING
0.24	WANTS	WANTED	0.18	CAPITAL	HAPPENED
0.24	HAPPEN	HAPPENED	0.18	OTHER	ANOTHER
0.24	FORCE	FORCES	0.17	AHEAD	AGAIN
0.23	HAPPENED	HAPPEN	0.16	WORKERS	WORDS
0.23	SERIOUS	SERIES	0.16	MEDIA	MEETING
0.23	TROOPS	GROUPS	0.16	UNITED	NIGHT
0.22	QUESTION	QUESTIONS	0.16	NEVER	SEVEN
0.21	PROBLEM	PROBABLY	0.15	WORLD	WORDS

Table 5. Most frequently confused word pairs.

Thank you.