The MAVA corpus

V. Aubanel, C. Davis, J. Kim
MARCS Institute for Brain, Behaviour and Development,
Western Sydney University

Last updated: 06 Oct. 2017

Contents

1 Short description **1**

2 Material and methods **1**
2.1 Sentence material . . . 1
2.2 Recording procedure . . . 1
2.3 Individual files extraction . . . 2
2.4 Audio and video post-processing . . . 2

3 List of files **3**
3.1 Corpus files . . . 3
3.2 Supplemental files . . . 3
3.2.1 Mosaics . . . 3
3.2.2 Metadata . . . 3

1 Short description

The MAVA corpus (MARCS Auditory-Visual Australian recordings of IEEE sentences) is a collection of high quality audiovisual recordings of 205 phonetically balanced sentences from the IEEE sentence database, recorded by a native Australian English female talker. The audio channel is annotated at the word and phoneme level. In addition, for the video channel, frame-by-frame lip contour X Y coordinates are provided. The center of the lip region is used as a reference for deriving four video regions: full face, upper face, lower face and lips. All files are freely available for download under the Creative Commons BY-NC-SA licence.

2 Material and methods

2.1 Sentence material

Sentences were selected from the IEEE sentences material Rothauser et al. (1969). All 720 sentences from the original material were run through a phonetic balancing procedure described in Aubanel et al. (2014), resulting in 72 lists of 10 phonetically balanced sentences. Phonetic balancing was run on keywords only and the phonemic transcription was obtained using the CMUdict [1] pronunciation dictionary for American English. The first 20 lists of 10 sentences were selected for the current material, with an additional 5 sentences included as supplement (to be used, e.g. for practice trials). Figure 1 shows the phonetic balance of the resulting selection as the average phoneme count for all keyword across the 20 lists of 10 sentences.

2.2 Recording procedure

A female speaker in her mid-twenties at time of recording produced the 205 sentences following a prompt presented on a computer screen. A light coloured sheet was used as the background for video capture. The video channel was recorded with a Sony PMW-EX1 camera with a resolution of 1920 by 1080 pixels

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict. Last checked on October 10, 2016

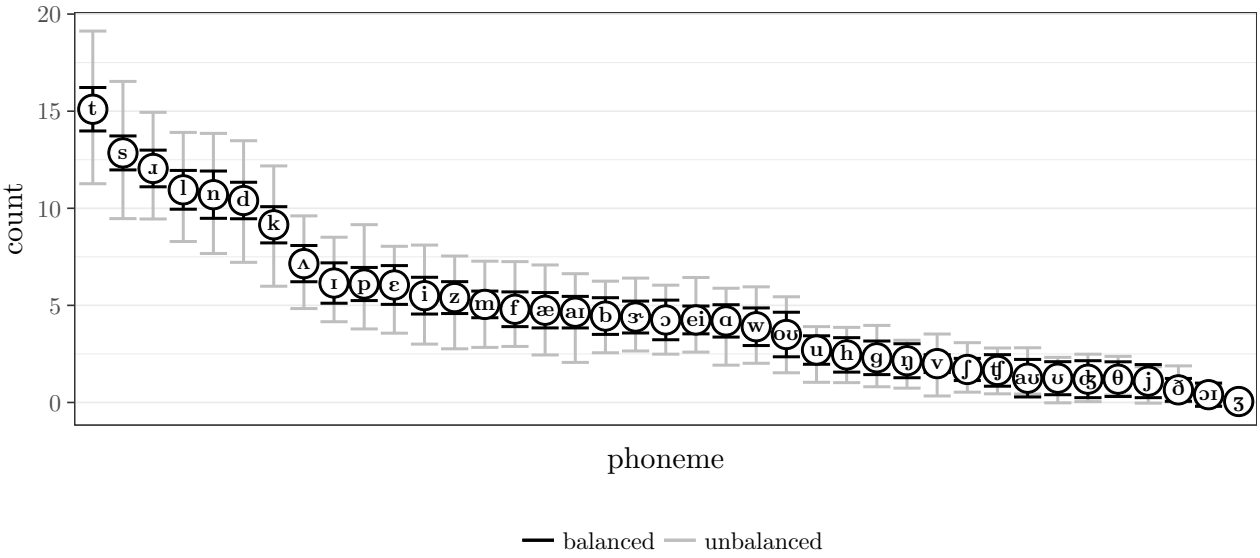


Figure 1. Average phoneme count for all keywords across 10-sentence lists. Black errorbars show ± 1 SD around the mean for the balanced 200-sentence corpus. Gray errorbars show ± 1 SD around the mean for the initial unbalanced 720-sentence IEEE corpus. There is a general agreement of the mean count values across phonemes for the initial unbalanced and balanced sets, although clear deviations are visible, e.g. for /f/ and /aɪ/. Note, other contiguous 200-sentence selections from the initial unbalanced 720-sentence set result in similar standard deviation as the initial set and more diverging mean count values (see Aubanel et al. (2014) for more details on the balancing procedure).

and a framerate of 50 frames per second, and later compressed using the video codec MPEG-4 H264-high. The audio channel was recorded with a Røde NTG-3 shotgun microphone placed at a distance of approximately 1.5 meter of the talker's mouth, and digitized with a MOTU UltraLite mk3 digital audio interface at 48 kHz sampling rate. Audio was also recorded with the built-in camera microphone and later used for synchronisation.

2.3 Individual files extraction

Sentence endpoints were semi-automatically identified and manually checked in the auditory signal, and served as a basis for individual audio and video files segmentation. The audio files were constructed by extracting a signal portion corresponding to the sentence duration with and additional 500 ms before the start and after the end of the sentence. Video files were constructed by extracting video frames corresponding to the audio interval, starting at the frame immediately preceding the audio onset. This results in a positive offset in audio files in relation to video files, uniformly distributed over the sentences, with a value in the interval $[0, 20]$ ms, i.e. lower than the duration of a video frame. This offset value is provided along each sentence of the corpus (see Section 3.2).

2.4 Audio and video post-processing

The audio channel was forced-aligned into word and phonemes and was manually checked and corrected. The lip contour was tracked using Sensarera (Bertolino, 2012), by manually initialising an region corresponding to the lip area and monitoring for departures away from the lip region in a frame-by-frame manner. Such departures were corrected by manually reinitialising the area a frew frames back and restarting the process. Once the lip contour tracking was completed, left (L) and right (R) lip corner coordinates were determined as the cartesian coordinates of the minimum and maximum abscissa coordinates of the lip contour respectively, and the lips midpoint as the coordinates of the midpoint of the segment linking L and R. Finally, the temporal average of the lips midpoint was defined as the mean coordinates of the lips midpoint across video frames spanning the duration of the sentence. Figure 2 shows these reference points for an example sentence. The temporal average of the lip's midpoint served as the origin for cropping the video channel into four different regions: full-face, upper-face, lower face and lips. The coordinate of the reference point is found in

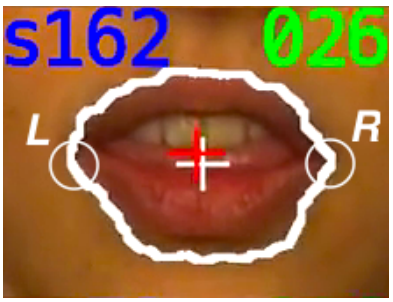


Figure 2. Lip region of sentence *s*162 at frame 26 (approx. 500 ms, corresponding to the onset of the sentence in the auditory channel). White contour: lip contour. White crosshair: lips midpoint, halfway between left and right lipcorners. Red crosshair: temporal average of lips midpoint, across the duration of the sentence. Extract from a region of mosaic `mosaic_lips_4KUHD_500ms.png` (see section 3.2).

the metadata (see section 3.2.2).

3 List of files

3.1 Corpus files

Table 1 lists the files available for each sentence of the corpus. Instructions for downloading the corpus are available at the MAVA main page: **http://dx.doi.org/10.4227/139/59a4c21a896a3**.

3.2 Supplemental files

In addition to sentence files, the following list of supplemental files are available for download.

3.2.1 Mosaics

For each video size (full face, upper face, lower face, lips), a mosaic of all 205 sentences was created allowing the corpus to be visualised at a glance. Mosaic videos include the sentence number and the frame number, the lip countour, the lips midpoint and the temporal average lips midpoint. Each mosaic comes in two spatial resolutions: 1080 lines (Full HD) and 2160 lines (4KUHD), and a static screenshot at 500 ms is provided. Mosaic files are available for download in the Attachment section of MAVA main page.

3.2.2 Metadata

A text file with relevant metadata is available for download. The fields are described in Table 2 and are also available individually for each item on the Alveo platform. The metadata file is available for download in the Attachment section of MAVA main page.

[2]http://www.fon.hum.uva.nl/praat/. Last checked on October 10, 2016

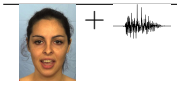
Illustration	Suffix	Description
	`_face_audio.mp4`	**Full-face audio-video mixture**. The size of the video is 672×896 pixels, referenced to the temporal average of the lips midpoint coordinates at $x = 336, y = 224$ (origin: bottom left). The audio channel is compressed with the AAC codec at a sampling frequency of 48 kHz and a bitrate of 192 kbit/s. No correction for the audio-video offset has been implemented, as it is deemed unnoticeable for this range of asynchronies below 20 ms. The audio-video offset value is nevertheless available, see section 3.2.2.
	`_face.mp4`	**Full-face video**. The size of the video is 672×896 pixels, referenced to the temporal average of the lips midpoint coordinates at $x = 336, y = 224$).
	`_faceup.mp4`	**Upper-face video**. The size of the video is 672×562 pixels, referenced to the temporal average of the lips midpoint coordinates at $x = 336, y = -110$. This reference point is outside the frame of the video, 110 pixels below the bottom border of the video.
	`_facelow.mp4`	**Lower-face video**. The size of the video is 672×334 pixels, referenced to the temporal average of the lips midpoint coordinates at $x = 336, y = 224$. Vertical concatenation of upper-face and lower-face videos would be equivalent to full-face videos.
	`_lips.mp4`	**Lips region video**. The size of the video is 256×192 pixels, referenced to the temporal average of the lips midpoint coordinates at $x = 128, y = 96$.
	`_16k.wav`	**16 kHz audio**. The sampling rate is 16 kHz.
	`_48k.wav`	**48 kHz audio**. The sampling rate is 48 kHz.
	`_annot.TextGrid`	**Annotation at the word and phoneme level**. Valid for both 16 kHz and 48 kHz sampling rates. The TextGrid format can be opened with PRAAT[2].
	`_lipcontour.txt`	**Lip contour**. Frame-by-frame coordinates of the lip contour. Coordinates are with reference to the original 1920×1080 pixels video dimensions.
	`_lipmidpoint.txt`	**Lip midpoint**. Frame-by-frame coordinates of the lip contour. Coordinates are with reference to the original 1920×1080 pixels video dimensions.

Table 1. Files available for each sentence. A thumbnail illustation of the content of the file is shown in the left column. The filename suffix is shown in the middle columnn and the detailed description in the right column.

Field	Description
`Identifier`	The sentence identifier.
`av_sync_spl`	The audio-video offset in samples, at 48 kHz sampling frequency.
`sync_s`	The audio-video offset in seconds.
ref_x	The x-coordinate of the temporal average lips midpoint used for reference in video files (see Table 1).
`ref_y`	The x-coordinate of the temporal average lips midpoint used for reference in video files (see Table 1).
`list_IEEE`	The list number of the sentence in the initial unbalanced 720-sentence set (Rothauser et al., 1969).
`item_IEEE`	The sentence number in the list of the unbalanced 720-sentence set (Rothauser et al., 1969).
`prompt`	The orthographic form of the sentence as read by the speaker in producing the sentence.

Table 2. Metadata fields available for each sentence.

References

Aubanel, V., García Lecumberri, M. L., and Cooke, M. (2014). The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology . *Int. J. Audiol.*, 53:633–638.

Bertolino, P. (2012). Sensarea: An authoring tool to create accurate clickable videos. In *CBMI*, Annecy, France.

Rothauser, E. H., Chapman, W. D., Guttman, N., Hecker, M. H. L., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., Weistock, M., McGee, V. E., Pachl, U. P., and Voiers, W. D. (1969). IEEE Recommended practice for speech quality measurements. *IEEE Trans. Audio Acoust.*, pages 225–246.