International Journal of Speech Technology
https://doi.org/10.1007/s10772-023-10030-3

Mouth2Audio: intelligible audio synthesis from videos with distinctive vowel articulation

Saurabh Garg[1,2] · Haoyao Ruan[1] · Ghassan Hamarneh[2] · Dawn M. Behne[3] · Allard Jongman[4] · Joan Sereno[4] · Yue Wang[1]

Received: 6 October 2022 / Accepted: 28 April 2023

Abstract
Humans use both auditory and facial cues to perceive speech, especially when auditory input is degraded, indicating a direct association between visual articulatory and acoustic speech information. This study investigates how well an audio signal of a word can be synthesized based on visual speech cues. Specifically, we synthesized audio waveforms of the vowels in monosyllabic English words from motion trajectories extracted from image sequences in the video recordings of the same words. The articulatory movements were recorded in two different speech styles: plain and clear. We designed a deep network trained on mouth landmark motion trajectories on a spectrogram and formant-based custom loss for different speech styles separately. Human and automatic evaluation show that our framework using visual cues can generate identifiable audio of the target vowels from distinct mouth landmark movements. Our results further demonstrate that intelligible audio can be synthesized from novel unseen talkers that were independent of the training data.

Keywords Audio synthesis · Mouth motion · Deep network · Vowel articulation · Speech intelligibility

1 Introduction

When we engage in face-to-face conversations, facial movements and corresponding voice are simultaneously used to perceive speech (Jongman et al., 2003; Kawase et al., 2015; Munhall et al., 2004).With multimedia (e.g., during video conferencing), we rely on visual facial cues when the audio is not transmitted well. Moreover, in noisy environments (e.g., cafeteria), seeing a talker's facial movements can particularly aid speech perception (Bernstein et al., 2004; Sumby & Pollack, 1954). One subsequent question is whether a missing or degraded audio signal can be recreated based on visual speech information extracted from a talker's face. Also, using visual speech to recreate an audio signal, perceptual accuracy and confounds can contribute to knowledge about the speech information potentially available from a visual signal and how it compares with an audio signal. Exploring such cross-modal synthesis of speech will not only contribute to our understanding of the interplay between the audible and visual components in speech communication but may have practical applications for the development of multimedia, multi-modal speech synthesis, as well as human–computer interface.[1]

To tackle this question, in the present study we leverage audio–video footage of talkers to model the relationships between the audio and video using deep learning approaches. We aim to develop an automated (video-to-audio) lip-reading system that can reconstruct the acoustic attributes of the talker's voice based on extracted visual speech attributes from the talker's facial movements. To this end, we selected a set of words containing representative English vowels that involve distinct visible facial articulatory movements that can be captured by a video camera. We then conduct audio synthesis of each vowel based on the corresponding facial

✉ Saurabh Garg
srbh.garg@gmail.com

1 Language and Brain Lab, Department of Linguistics, Simon Fraser University, Burnaby, Canada

2 School of Computing Science, Simon Fraser University, Burnaby, Canada

3 Department of Psychology, Norwegian University of Science and Technology, Trondheim, Norway

4 KU Phonetics and Psycholinguistics Lab, Department of Linguistics, The University of Kansas, Lawrence, USA

[1] https://github.com/srbhgarg/VowSynth.git.

Published online: 25 May 2023

articulatory movements, taking into account variances due to speaking style and talker gender, as well as tensity of articulatory gestures.

1.1 Articulatory and acoustic correlates of vowels

Given that acoustic variations in speech are triggered by alterations in articulatory configurations, such articulatory variations are conceivably measurable and can be perceived to aid intelligibility. Indeed, kinematic measurements showed positive correlations among articulation, acoustics, and intelligibility measures of speech segments including vowels (Gagné et al., 2002; Kim & Davis, 2014; Tasko & Greilick, 2010). Specifically, the articulation of different vowels is characterized by distinct facial movements, for example, horizontal lip spreading for unrounded vowels such as /i, ɪ/, jaw lowering for low vowels such as /ɑ, ʌ/, and lip rounding for rounded vowels such as /u, ʊ/ (Kim & Davis, 2014; Tang et al., 2015; Tasko & Greilick, 2010). In addition, a greater extent of such movements has been observed for tense vowels (/i, ɑ, u/) compared to lax vowels (/ɪ, ʌ, ʊ/) (Tang et al., 2015), and for vowels spoken in a clear style versus plain, conversational speaking style (Kim & Davis, 2014; Tasko & Greilick, 2010). Further, these articulatory movements have been found to be visually salient and have been used as helpful sources of linguistic information for speech perception (Kim & Davis, 2014; Tasko & Greilick, 2010; Traunmüller & Öhrström, 2007).

These articulatory motion results are aligned with previous findings on the acoustic features of vowels. In particular, the more expanded acoustic vowel space and more peripheral formant frequencies found for clear speech (Bond & Moore, 1994; Bradlow et al., 1996; Ferguson & Kewley-Port, 2002, 2007) have been attributed to more extreme articulatory movements, including greater vertical lip movement, jaw lowering, horizontal lip stretching, and lip protrusion (Redmon et al., 2020; Tasko & Greilick, 2010). For example, acoustic studies show that the second formant of the front vowels (/i, ɪ/) generally increases in clear speech relative to plain speech (Ferguson & Kewley-Port, 2002; Ferguson & Quené, 2014; Lu & Cooke, 2008), which is conceivably due, in part, to the shortening of the vocal tract resulting from greater lip-spreading in clear speech. Correspondingly, the production of these front vowels involves horizontal lip spreading, and articulatory studies show they have greater horizontal lip movement (Tang et al., 2015). Likewise, the greater degree of lip rounding and lip protrusion for the rounded vowels /u, ʊ/ in clear versus plain speech (Tang et al., 2015) results in vocal tract lengthening which consequently lowers the second formant of these rounded vowels in clear speech (Ferguson & Kewley-Port, 2002). In terms of duration, articulatory studies reveal an overall greater and longer articulatory movement for tense vowels compared to lax vowels (Tang et al., 2015). Acoustically, tense vowels are typically longer than their lax vowel counterparts, presumably resulting from the longer excursions for the articulators to reach the more extreme tense vowel target positions (Hillenbrand et al., 1995; Watson & Harrington, 1999). Furthermore, previous studies by Tang et al. (2015) and Leung et al. (2016) have shown talker gender interaction with vowel tensity and speech style. Tang et al. (2015) show that male compared to female talkers have larger clear-plain distinctions in visual articulatory movements, whereas Leung et al. (2016) show that the acoustic patterns of clear speech modifications do not differ between male and female talkers.

These results convincingly indicate a direct relationship between visible articulatory movements and acoustic characteristics of vowels across speech styles and vowel tensity, which provides the foundation for the current cross-modal synthesis study.

1.2 Automatic video-to-audio speech synthesis

While text-to-speech synthesis has matured, cross-modal audio synthesis based on articulatory information is still developing and faces many challenges.

Research on automatic visual to auditory speech synthesis can be broadly classified into two approaches: the first approach, termed “silent speech interfaces” (SSI), relates to generating audio from biosignals or ultrasound videos of tongue movements (Freitas et al., 2017; Gonzalez-Lopez et al., 2020), while the second approach maps facial movements to audio or spectrogram directly (Yehia et al., 1998; Akbari et al., 2018; Ephrat & Peleg 2017; Vougioukas et al., 2019).

Most SSI studies rely on different biosignals to communicate, such as electrophysiological recordings of neural activity (Anumanchipalli et al. 2018; Herff et al., 2015), electromyographic (EMG) recordings of vocal tract movements (Schultz & Wand, 2010) or the direct tracking of articulator movements using imaging techniques (Hueber et al., 2010). These techniques rely on non-acoustic signals that are generated during speech production to restore audio. Of these, EMG, permanent magnetic articulography and electromagnetic articulography are most commonly used and involve placing markers/sensors on the body (Schultz & Wand, 2010). A mapping function is then learnt using machine learning techniques that map the recorded biosignal to the audio speech. However, the placement of sensors on the mouth and tongue may make the mouth/face movement unnatural and sensors are not present in natural speech settings. The present study uses a standard video or articulation approach that does not require sensors and intentionally avoids these potential limitations (Tang et al., 2015). Secondly, the video-based methods that do exist in SSI use

Table 1 Comparison of the approach in the current study to other existing research

Study	Method	Output	Video encoder	Audio encoder	Phase accounted for?	Unseen talkers?	Human assessment
Wang et al. (2022)	VCVTS	Mel spectrogram	Con-vT3D + ResNet-18 + temporal CNN-4	VQCPC	Yes	Yes	No
Saleem et al. (2022)	E2E-V2SResNet	Spectrogram	CNN + ResNet-18	ConvT-6	Yes	No	No
Prajwal et al. (2020)	Lip2wav	Audio	CNN + LSTM	Tacotron2	Yes	No	Yes
Mira et al. (2022)	GAN-based	Audio	ResNet-18 + biGRU	Conv1D-GAN	No	Yes	No
Akbari et al. (2018)	Lip2AudSpec	Spectrogram	CNN-7 + LSTM	Dense-2	Yes	No	Yes
Ephrat andPeleg (2017)	Vid2Speech	Audio	CNN-5	CNN-5	No	No	Yes
Le Cornu and Milner (2015)	DNN_UNV	Audio	Dense-3	STRAIGHT	No	No	Yes
Assael et al. (2016)	LipNet	Text	STCNN3 + BiGRU	N/A	N/A	Yes	Yes
Current study	Mouth2Audio	Audio	Landmark	ConvT1D-5	Yes	Yes	Yes

ConvT1D denotes a network built of 1-dimensional convolutional transpose layers where the number of layers used by each network is shown after each hyphen."Unseen talkers?" specifies whether evaluation included novel unseen talkers ("yes") or the same talkers in both model training and testing ("no")

ultrasound images of the tongue along with lip movements to synthesize audio (Hueber et al., 2010). In these methods, an ultrasound machine is placed under the speaker's chin and the tongue movement is tracked. Machine learning methods such as hidden Markov model (HMM) or deep neural networks are then trained to learn the mapping from the ultrasound signal to the audio (Hueber et al., 2010). Although tongue movements are directly related to speech articulation and the resultant acoustic signal, extracting visible facial articulatory cues is more accurate (Yehia et al., 1998) and practical for applications in face-to-face communication, as well as video-based communication.

The second type of approach for automatic visual to auditory speech synthesis uses the video of the talker's face to learn the mapping of the movement of the lips and lower face to the text or the audio (Assael et al. 2016; Saleem et al., 2022; Vougioukas et al., 2019; Wang et al., 2022). With the current advances in deep learning, these methods use stacked layers of neural network to learn the mappings. For example, Saleem et al. (2022) proposed a deep convolutional encoder-decoder framework (E2E-ResNet) that captures face video and encodes video frames to a latent space using ResNet and then decodes the latent representation into a corresponding spectrogram. Wang et al. (2022) proposed deep net based on vector quantization with contrastive predictive coding for content encoder to learn discrete acoustic units and a multi-layer Lip-to-index network to learn the mapping of lips to the indices of the above learned discrete acoustic units, called Voice-Conversion-Video-to-speech (VCVTS). Assael et al. (2016) proposed a deep model that uses spatiotemporal convolutions and gated recurrent nets to learn the mapping of video frames of the speaker's mouth to the sentence-level text (LipNet). Ephrat and Peleg (2017) trained a deep network consisting of a convolutional neural network (CNN) to learn the mapping between video of lips and the linear predictive coding (LPC) features of the corresponding audio which are then used to reconstruct the audio (Vid2Speech). The method Lip2AudSpec proposed by Akbari et al. (2018) also used two deep neural networks: one autoencoder network to encode/decode a spectrogram and another 7-layer 3D CNN and long short term memory (LSTM) to learn mapping from video to the encoded audio spectrogram as learnt by the first autoencoder network. Most studies on audio synthesis (e.g. Akbari et al., 2018; Assael et al., 2016; Le Cornu et al., 2015; Mira et al., 2022; Saleem et al., 2022; Vougioukas et al., 2019; Wang et al., 2022) employed the GRID corpus (Cooke et al., 2006) that consists of many hours (approx. 50 min of speech per talker) of video clips of 6-word sentences presented in a fixed order of [command]-[color]-[preposition]-[letter]-[digit]-[adverb], where each [position] has 4-word choices except that [letter] and [digit] have 25- and 10-word choices, respectively. Although the corpus has a vocabulary of 51 words, the words are acoustically very distinct, consisting of different vowels and consonants. The performance of these studies was evaluated using word error rate (WER) which was generally reported to range from 40 to 50%, whereas Mira et al. (2022) reported a WER of 23.13%. Their problem formulation using sentences benefits from the ability to encode temporal priors, which would not be feasible for word-level synthesis. Further, as summarized in Table 1, aside from only one method

(Akbari et al., 2018), phase information of the synthesized audio is ignored in all aforementioned research, thus leading to poor realism, as demonstrated, for instance, by the published audio clips of Vougioukas et al. (2019). In contrast, we made conscious design decisions to address phase information as will be elaborated in Sect. 3.3. As seen from Table 1, previously, only Assael et al. (2016) and Vougioukas et al. (2019) have evaluated performance on unseen talkers, where Assael et al. (2016) produces text as output instead of audio, and Mira et al. (2022) lacks any human assessment.

1.3 The present study

Different from previous studies (Akbari et al., 2018; Mira et al., 2022; Vougioukas et al., 2019) where the target words differ in both consonants and vowels that typically appear in sentential contexts, the current study focused on the more challenging case of isolated target words that are minimally contrastive. As motivated previously in Tang et al., 2015, we selected monosyllabic English words that differ only in their vowels: keyed, kid, cod, cud, cooed, and could. These six vowels are so chosen because they comprehensively span the vowel space (Tang et al., 2015). More specifically, the six English vowels in these words, /i, ɪ, ɑ, ʌ, u, ʊ/, contain visible articulatory movements of lip spreading, jaw lowering, and lip rounding and protrusion (Tang et al., 2015). These six vowels also form three tense-lax vowel pairs, with the tense vowels /i, ɑ, u/ having more extreme articulatory movements and longer duration and thus higher visual salience than the lax vowels /ɪ, ʌ, ʊ/ (Tang et al., 2015). In addition, our dataset contains productions of these words in clear as well as plain, conversational speaking styles produced by multiple male and female talkers to take into account variability in each of the vowel categories. Given this span and variability in these vowels, we believe that our approach not only will generalize to other vowels, but also complement the existing literature that largely does not address minimally contrastive words.

Based on the correspondence between visual articulatory and acoustic features of these vowels reviewed above, we hypothesize that the acoustic consequences of such articulatory gestures should be retrievable. This dataset makes our study unique compared to aforementioned studies (Akbari et al., 2018; Mira et al., 2022; Vougioukas et al., 2019) in that (1) focusing on the most salient component in a word, i.e., the vowel, enables extraction of the most distinct features (i.e., steady-state formant patterns corresponding to the vowels) that are not obscured by variances due to coarticulatory effects from different adjacent segments; (2) using isolated words rather than sentences allows identification of the words without the possibility to rely on contextual cues in a sentence and (3) the first two points, along with the inclusion of vowel tensity, style and talker variations, make our approach highly generalizable.

The problem we take on is especially challenging in four ways. Unlike common lip-reading frameworks (e.g., Assael et al., 2016) that generate low-dimensional outputs with word-classification typically being the desired final output, our problem formulation requires our outputs to be high-dimensional audio waveforms that are intelligible. Second, our dataset not only contains word tokens with distinctive articulatory features (e.g., /i/ vs /u/), but also those with similar articulatory cues differing only in tensity (e.g., /i/ vs /ɪ/). Third, the performance of our automated video-to-audio system is evaluated on unseen talkers that were not part of the training set. Also, the evaluation was done not only using ASR to allow a broad comparison with previous related studies, but also using human perceivers so that the use of our automated video-to-audio system can be extended to real-life applications. Lastly, the system's performance was evaluated on two different speech styles: plain and clear.

To overcome the aforementioned challenges, we designed a network that strictly takes only mouth motion data as input and outputs an audio waveform that best corresponds to the motion captured by the video data, that is, using direct mapping, as opposed to encoding mouth movements in video (Akbari et al., 2018; Vougioukas et al., 2019). We adopted a deep learning approach in which the training objective is to minimize the error between the spectrograms of the generated audio waveform and those of the actual waveforms. To the best of our knowledge, our study is the first to develop a video-to-audio framework discriminating between very similar words that vary only in their vowel, is robust to different speech styles and uses formant-based learning (i.e., training based on formant frequencies that characterize individual vowels) for improving audio intelligibility, and thus shows that motion trajectories can be used to synthesize corresponding audio.

To summarize, our approach integrates computational approaches with phonetic insight, which has not been systematically adopted by either field. Based on knowledge of articulatory-acoustic correspondence, we expect that machine recognition of a specific set of articulatory attributes characterizing a certain sound can lead to reconstruction of the acoustic information of this sound. Using individual attributes as building blocks allows our system to circumvent context- and speaker-induced speech variance, thus making the output more generalizable.

2 Materials

Our problem formulation involves the use of a carefully curated dataset created in previous work (Tang et al., 2015). Specifically, three English vowel pairs, /i-ɪ/, /ɑ-ʌ/, /u-ʊ/ differing in articulatory features were the target vowels for

examination, with the tense vowels /i, ɑ, u / having more extreme articulatory movements than the lax vowels /ɪ, ʌ, ʊ/. These English tense and lax vowels were embedded in monosyllabic /kVd/ contexts, resulting in six common English words: “keyed”, “kid”, “cod”, “cud”, “cooed”, and “could”. Unlike previous studies such as those summarized in Table 1, where the target words differ in both consonants and vowels, we use carefully chosen minimally contrastive target words which include the same consonants but differ only in the vowel. Using minimally contrastive words enables extraction of the most distinct features for the /kVd/ syllable without being obscured by variances due to coarticulatory effects in different consonantal contexts. This thus allows direct correlation between the articulatory movements and the acoustic cues of particular segments.

As revealed in Tang et al. (2015), the vowels used here involve visible and measurable articulatory differences, (e.g., greater horizontal lip movements for “keyed, kid”, greater vertical lip movements and jaw lowering for “cod, cud”, and greater lip rounding for “cooed, could”). The selected vowels also differ in tensity, that is, how “extreme” articulatory movements may be (tense vowels—more extreme, as in “keyed, cod, cooed”, and lax vowels—less extreme, as in “kid, cud, could”). Furthermore, the production of each token was recorded in isolation (as opposed to continuous speech) and articulated in two speech styles: plain (conversational) and clear (more enunciated). The terms “plain (conversational) speech” and “clear (more enunciated) speech” are used based on the convention used in previous clear-speech studies (e.g., Ferguson & Kewley-Port, 2002; Maniwa et al., 2008) including our own (Leung et al., 2016; Redmon et al., 2020; Tang et al., 2015). These two terms refer to the contrasting speech styles resulting from instructions to talkers to speak a word or an utterance “naturally” first in the manner used in a plain, natural conversation, and then repeat it “clearly” in order to improve intelligibility (See the detailed information below on the elicitation of plain and clear speech stimuli used in this study).

Akbari et al. (2018) mention that although many consonants can be recovered from lip movements, reconstructing different vowels accurately is quite difficult. Compared to previous work (e.g., Akbari et al., 2018) where multiple cues are available to characterize a word (e.g., consonants, vowels, semantic and contextual information), the words used here present a more challenging test which allows application of our approach to challenging communication contexts (e.g., noisy environments) where similar words are difficult to distinguish.

Audio–video recordings of the target words were obtained from eighteen (eight male and ten female) talkers. The talkers (aged 17–30, mean: 22) were recruited from the student population at Simon Fraser University (SFU). They reported no history of speech or language impairment. Audio–video recordings were acquired in a sound-attenuated booth in the Language and Brain Lab at SFU. Front-view videos were captured with a Canon Vixia HF30 camera at a recording rate of 29 frames/second. Audio recordings were acquired simultaneously using Sonic Foundry Sound Forge 6.4 at a sampling rate of 48 kHz, with a Shure KSM microphone placed at a 45-degree angle, 20 cm away from the talker’s mouth.

During recording, a talker articulated each of the 6 words multiple times in a random order for each of the two speech styles. The plain (conversational) and clear (more enunciated) styles were elicited using a simulated interactive computer speech recognition program established previously (Maniwa et al., 2009; Redmon et al., 2020). The computer instructed the talker to pronounce each token displayed on the screen naturally to generate plain-style productions. Then, the software would display the program's identification of the spoken token, involving erroneous "guesses". If a guess was incorrect, the talker would be asked to repeat the token more clearly to facilitate the software's ability to distinguish the confused token and thus generate a clear-style production. For correct guesses, no clear production was required. Thus, with repetitions, for each word and each talker, fifteen plain productions and twelve clear productions were elicited, resulting in a total of 2916 audio–video words [plain (6 words × 18 talkers × 15 repetitions) + clear (6 words × 18 talkers × 12 repetitions)], all of which had been evaluated as correct productions of the target words by two native Canadian English speakers.

3 Video-to-audio synthesis

Several intermediate steps are involved in estimating the audio from the articulatory movements of the face. In order to capture the articulatory movements, the face needs to be located in the video and various facial landmark points must be identified and tracked over the course of the video. These landmark points are then used to train a machine-learning model that learns the mapping from the landmark trajectory to the audio by minimizing spectrogram-based and formant-based loss function. More details on the individual steps are provided in the following section.

3.1 Preprocessing of training and test data

3.1.1 Video Segmentation

All tokens from each talker were recorded in a single video file and thus had to be automatically segmented using features extracted from audio. Specifically, energy was computed from the audio signal and a threshold was

empirically estimated to separate the silence and the audio signal. The segmented audio signal timings were then transferred over to the video signal and segmented into separate tokens. More information on this step can be found in our previous work (Garg et al., 2019).

3.1.2 Face detection

In order to extract features from faces in the videos, the first step is to detect the location of the face in the image. In our experiment setup each video has only one face and the face is always facing the camera. To detect the frontal face, two different methods were tried: Multi-Task Convolutional Neural Network (MTCNN) (Zhang et al., 2016) and dlib python library by (King et al., 2009). The choice of MTCNN over alternatives such as dlib was derived from preliminary experiments where we observed dlib to be less consistent than MTCNN in its location of the face bounding box. We employed MTCNN to detect the bounding box of the face shown on the first frame of each video token. The face was detected in a single frame of the video and the same coordinates were used for the rest of the video. Since our tokens are no longer than 2 s, we did not observe any large head movement during the utterance for the face to move out of the reference bounding box. The face detection step is important since if the bounding box misses a part of the face, the results of the next step are drastically affected. A buffer with a fixed height (100 pixels) was also added around the detected box.

3.1.3 Face landmark detection

Face landmark detection was performed on the detected face. We used a convolution neural network—conditional random field (CNN-CRF) as proposed in Chen et al. (2019) for a face landmark detector that is trained to detect 68 facial landmarks. The method jointly trains CNN and CRF, where CNN is used to detect landmark points and CRF encoded the structural relationship between the different landmark points. In the current study only landmarks from the lips and lower jaw were used to train our network, leading to inputs of motion trajectories from k = 29 landmarks.

3.1.4 Face landmark tracking

The landmark points are detected on each of the f video frames ($f = 80$) separately to obtain the coordinates of each landmark point $\mathbf{x}$ over time. The choice of 80 frames was based on the longest token that we had in our dataset. The shorter tokens were padded on either end with empty frames to make them 80 frames long. The obtained points are smoothed using a Savitzky–Golay filter (Savitzky & Golay, 1964) of order 4 to remove any jitter present in the data using the Python scipy-signal library (savgol filter). The filter replaces each sample by fitting a polynomial on $2n+1$ neighboring points where n is greater than the order of the polynomial. Based on empirical evaluation, we set $n = 5$.

3.1.5 Resampling

The number of frames over which each landmark is tracked can vary across video tokens, so they are resampled to have the same size using cubic interpolation.

The above resampled data is used to train a deep network that takes trajectories of detected 29 lip and jaw landmark points during the token utterance as input and that outputs the corresponding audio waveform. The details of the network are provided in the following section.

3.2 Formant network

Formant frequencies, particularly the first three formants (F1, F2 and F3) have been found to be the main acoustic correlates to vowel production and identification (e.g., Hillenbrand et al., 1995; Peterson & Barney, 1952). In the present study, we focus on generating the formant frequencies F1, F2, and F3 for each of the vowels, based on acoustic information. To compute the formant frequencies, we trained another deep network consisting of fully connected layers. This network was trained on the audio extracted from the same dataset as the vowel synthesis network. The network takes the spectrogram of the audio as input and outputs the first three formant frequencies and their corresponding three bandwidths (B1, B2, B3) for each token for each talker as shown in Fig. 1. The ground truth formant frequencies and their bandwidths were obtained using PRAAT toolbox (Boersma, 2001). The network is trained until the difference between the predicted output (3 formant frequencies and bandwidths) and the corresponding ground truth values have reached a minimum. This proposed network consists of three time-distributed fully connected layers to make predictions for every time window of the spectrogram. The first layer contains 200 nodes followed by 128 and the final layer has six nodes for each of the six outputs (three formant frequencies and their corresponding bandwidths). The mean absolute error on test tokens was 88.94 Hz, 198.54 Hz, 255.49 Hz for F1, F2 and F3, respectively, over all token types. Table 2 summarizes the mean absolute error on each token separately. Further, the Pearson correlation analysis shows that a strong significant correlation between the estimated formant frequencies and the ground-truth frequencies: [F1: r(936) = 0.80); F2: r(936) = 0.84; F3: r(936) = 0.64; p-values < 0.0001].These results indicate that the network was able to estimate the formant frequencies with reasonable accuracy.

A similar analysis was also performed for bandwidth. The mean absolute error on test tokens was 139.05 Hz, 263.71 Hz and 346.79 H for B1, B2 and B3, respectively, over all token

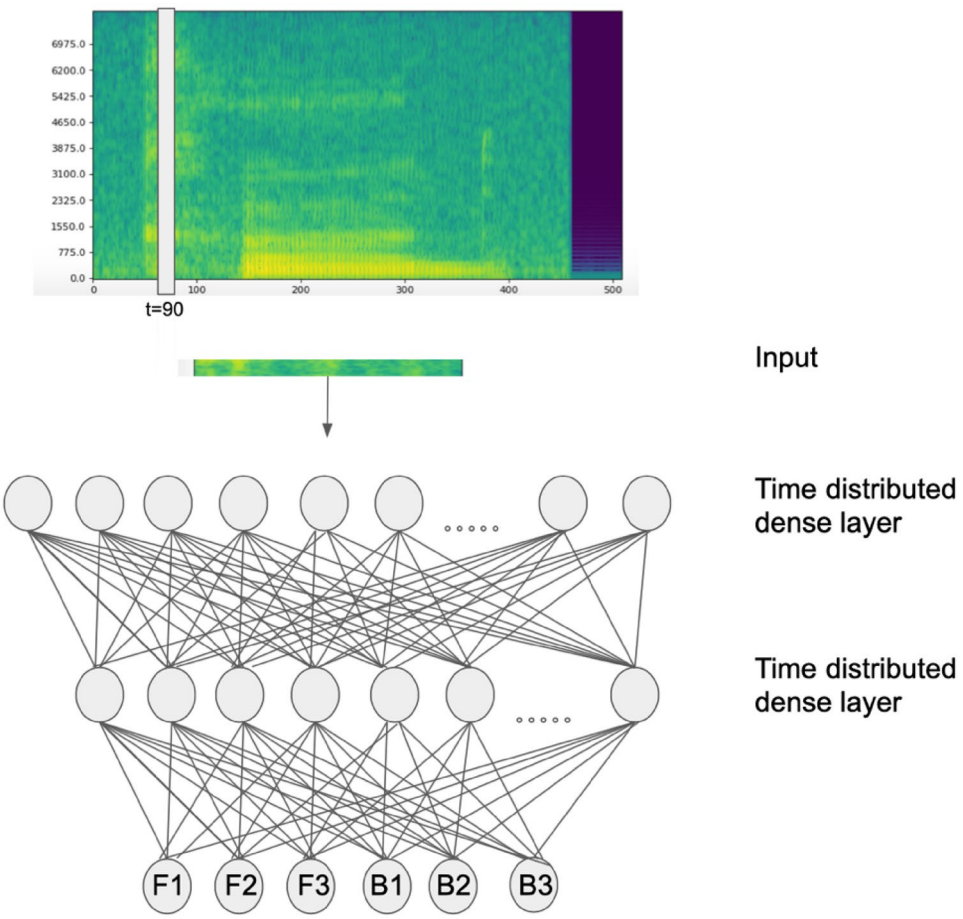


Fig. 1 Formant estimation network. The network takes a column of spectrogram as input and estimates formant frequencies (F1, F2, F3) and their bandwidths (B1, B2, B3) of that part of the spectrogram

Table 2 Mean absolute error between the predicted formant frequencies (F1, F2, F3) by the trained formant network of each token from the PRAAT generated formant frequencies (ground-truth) of the same token

Token type	F1 (in Hz)	F2 (in Hz)	F3 (in Hz)
keyed	− 60.35	− 262.18	− 253.36
kid	+ 93.75	− 204.34	− 282.22
cod	+ 87.68	+ 144.05	+ 220.46
cud	+ 86.25	+ 161.16	− 237.25
cooed	− 88.69	− 228.52	− 316.60
could	+ 118.83	+ 193.42	+ 230.25

The direction of the difference was also measured separately and indicated by ± where "+" indicates the predicted measurement to be higher than ground-truth and "−" indicates the measurement to be smaller than the ground-truth

types. Further, Pearson correlation analysis showed a significant correlation between the predicted bandwidths by our network and the Praat estimated bandwidths for each of the bandwidth measures: [B1: r(936) = 0.50; B2: r(936) = 0.30; B3: r(936) = 0.25; p-values < 0.0001]. The lower magnitude of correlation coefficients for bandwidths compared to that for formant frequencies is presumably due to less accurate bandwidth estimates by Praat and other automated acoustic analysis systems as reported previously (Burris et al., 2014). For this reason, no in-depth analysis was performed.

3.3 Network loss function

As shown in Fig. 2, our deep network $g(x)$ consists of a dense layer and 5 1D transpose convolution (conv1d-transpose) layers. Each layer is followed by batch normalization and ReLu (rectified linear unit) nonlinearity except for the last layer which uses tanh nonlinearity. Our network then takes as input motion trajectories ($x \in R^{2k \times f}$) of k facial landmarks extracted from each word token consisting of f video frames and learns to output the corresponding audio (y_{gen}) of fixed length. It does so by minimizing the following loss function:

$$L_{total} = w_{spec} L_{spec} + w_{formant} L_{formant}$$

where L_{spec} is the loss term associated with spectrogram and $L_{formant}$ is the loss term associated with the formant frequencies. The larger the disagreement between the generated and the ground truth spectrogram, the larger is the loss value.

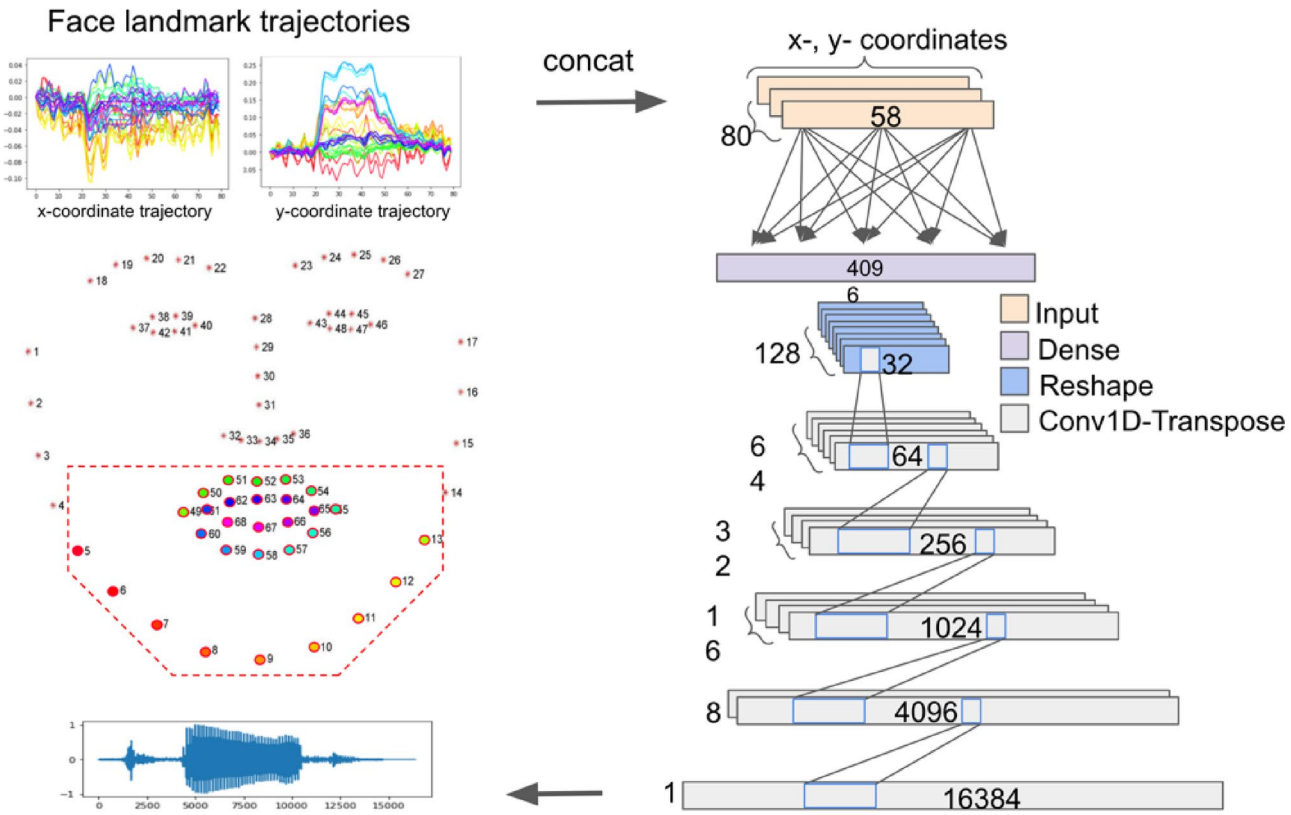


Fig. 2 Input and output (left) to our network (right)

w_{spec} and $w_{formant}$ are the respective scalar weights that control how much each of the above loss terms contributes to the total loss. w_{spec} and $w_{formant}$ were chosen empirically after repeated experiments that gave the best performance. In our experiments, w_{spec} was set to 1 and $w_{formant}$ was set to 10. L_{spec} is computed using the following equation:

$$L_{spec} = \frac{1}{N}\sum_{n=1}^{N}\left(\log\left(\left|F\left(y_{ngen};l,s\right)\right|\right) - \log\left(\left|F\left(y_{ngt};l,s\right)\right|\right)\right)2$$

where N is the batch size (set to 32); y_{gen} and y_{gt} are generated audio and ground truth audio waveforms, respectively; F denotes the short-time Fourier transform (STFT) that computes a spectrogram from an audio waveform using hyperparameters l and s, with l denoting the frame length ($l=128$), and s denoting the frame step ($s=32$ s). The loss term associated with formants: $L_{formant}$ is computed as follows:

$$L_{formant} = \frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{3}\left(w_i(Fmt_{igen} - Fmt_{igt})\right)^2 + \frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{3}\left(w_i(Bw_{igen} - Bw_{igt})\right)^2$$

where N is the batch size, $Fmt_{igen} and Fmt_{igt}$ are the i-th generated and ground truth formant frequencies, and $Bw_{igen} and Bw_{igt}$ are the corresponding i-th generated and ground truth bandwidths, respectively. Since the larger formant frequency values will have larger deviations, they will end up contributing more to the loss function, so the contribution of each formant frequency is normalized using fixed weights w_i. The weights w_i were empirically determined and set to 50, 100 and 200 for the three formant frequencies, respectively.

Network training is terminated by convergence of the loss computed on the validation set. Based on preliminary experiments, we decided to train our generator using Adam's optimizer with a learning rate of 0.01 and decay rate of 1e-5. Each Conv1d-transpose layer uses a kernel of size 25 and a stride of 4 except for the first layer which uses a stride of 2. In our study, the trained network architecture is the same for all the talkers[2] but the network was trained separately on each gender.

[2] To add talker-related information to the output speech, we could condition the network by feeding the speaker ID to the model in the form of one-hot encoding. At the time of testing, when a particular speaker's voice is to be generated, we could provide the corresponding speaker's one-hot encoding vector. This will be similar to what is already done in WaveNet (Oord et al., 2016).

3.4 Post-processing to incorporate phase

While the human ear is commonly assumed to be insensitive to audio phase, recent papers (e.g., Laitinen et al., 2013) suggest otherwise. Inspired by Laitinen et al. (2013), we conducted preliminary experiments and observed that when phase was reconstructed from spectrograms using the Griffin-Lim method (Griffin & Lim, 1984), the audio sounded unnatural. Further, we discovered that the perceptual quality of a generated token could be improved by using a 'template' phase. While experimenting with template phase, we realized that even information coming from another token from another talker still improves the perceptual quality of the audio. Based on these observations, we created a template phase by extracting the phase information from a randomly drawn audio token and used this template phase when reconstructing the new audio in the final reconstruction step, i.e.:

$$y_{gen}(t) = \frac{1}{2\pi} \int |Y(j\omega)| e^{\left(j\left(\omega t + \angle \widetilde{Y}(j\omega)\right)\right)} d\omega$$

where y represents the audio in the time-domain; Y represents the STFT of y; ω denotes the frequency; t denotes time; j represents complex number; $Y(j\omega)$ represents the magnitude of the spectrogram of the generated audio; and $\angle \widetilde{Y}(j\omega)$ represents the template phase added to the generated audio.

4 Audio synthesis evaluation

We evaluated our framework along two dimensions: standard quantitative methods that measure the intelligibility and quality of the generated audio data (Experiment 1), and human intelligibility testing (Experiment 2).

4.1 Experiment 1—evaluation by automatic speech recognition

4.1.1 Methods

For the automatic evaluation, the following four metrics were employed: the mean mel-cepstral distortion (MCD) that measures the distance between two signals in the mel-frequency cepstrum, which allows us to assess the perceptual difference between the generated audio data with respect to the real audio data; Short Term Objective Intelligibility (STOI) and Perceptual Evaluation of Speech Quality (PESQ) metrics to respectively measure the intelligibility and quality of the generated audio; and lastly, Word Error Rate (WER), which measures the rate at which an audio is mislabeled. Smaller values are preferred for MCD and WER whereas larger values are preferred for STOI and PESQ.

To measure WER, we used an approach similar to (Vougioukas et al., 2019) where an automatic speech recognition algorithm (ASR) is trained on the ground truth audio using a set of nine features. i.e., MFCC coefficients, chromogram, coefficients of fitting a 3rd-order polynomial on the spectrogram, spectral centroid, bandwidth, roll off, zero-crossing rate, tonnetz (Harte et al., 2006) and spectral flatness (Dubnov, 2004). The features were computed using a Python library (librosa). We trained seven different classifiers for word classification using well known classifiers (such as SVM, random forest) but, due to space limitations, report results only for the top two classifiers (based on WER), namely Random Forest and neural network consisting of one layer with 100 hidden sizes, both of which gave consistently high performance on the training set.

To be able to measure the performance of the trained network on novel unseen talkers, the full set of talkers was divided into two groups: training and testing talkers. Four talkers (2 female, 2 male) were randomly sampled from the whole dataset to be part of the test group. All the tokens obtained from these four talkers then formed the test set whereas tokens from the remaining 14 talkers formed the training set. Since the talkers from the test set are independent from the training set, performance of our video-to-audio synthesis system was evaluated on unseen talkers and tokens.

4.1.2 Results

The ASR's classification responses were then compared with the true labels to facilitate the calculation of WER. Using these performance metrics, we next conducted various experiments and highlight the key findings below.

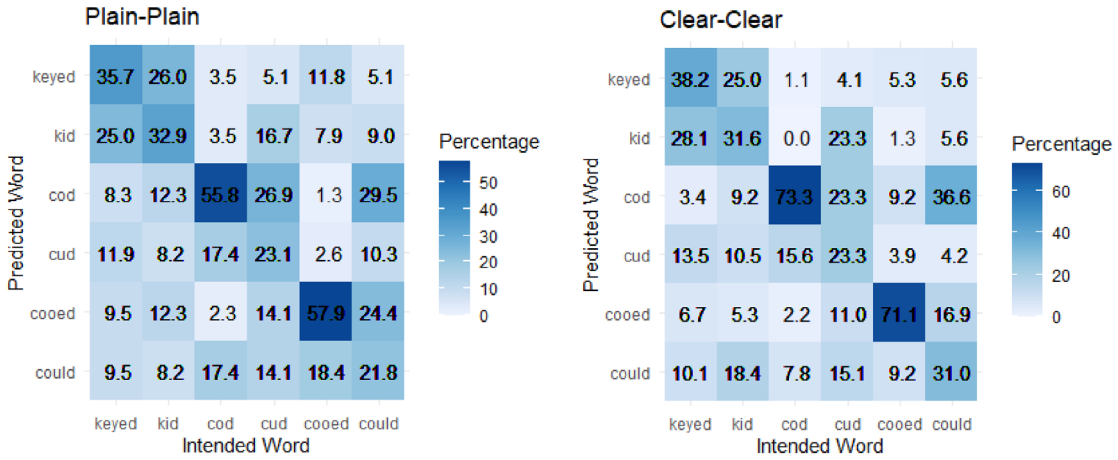
4.1.2.1 Phase reconstruction

Table 3 reports two approaches to handle phase: the Griffin-Lim algorithm and our proposed phase template method. Quantitatively, using Griffin-Lim's algorithm for phase reconstruction yielded slightly higher WER than our proposed approach using the phase template for reconstruction. Griffin-Lim also performed slightly worse than our proposed method in terms of increase in distortion (MCD) and decreases in evaluated intelligibility (STOI and PESQ). We tested our model using two different sampling rates, 16 kHz and 8 kHz. This helps to compare our method more fairly with different earlier methods that generated audio at different sampling rates. In general, generating audio at 16 kHz is a more complicated problem as compared to generating audio at 8 kHz since using the same input twice as many numbers need to be predicted at 16 kHz as compared to

Table 3 Performance evaluation of our method compared to the latest methods

Sampling rate of generated audio	Saleem et al., (2022)	Wang et al. (2022)	Mira et al., (2022)	Vougioukas et al., (2019)	Current method			
	Fast Griffin-Lim	Griffin-Lim			Fast Griffin-Lim		Griffin-Lim	
	16 kHz	Not reported	16 kHz	8 kHz	16 kHz	8 kHz	16 kHz	8 kHz
PESQ (↑)	2.03	1.417	1.47	1.24	1.34	1.42	1.37	1.46
STOI (↑)	0.627	0.582	0.523	0.445	0.594	0.594	0.598	0.598
MCD (↓)	27.79	8.36	37.91	24.30	5.06	6.78	5.03	6.51
WER (↓)	13.77%	Not reported	23.1%	40.5%	35.2%	–	34.4%	–
WER-h (↓)					N/A	–	38.1%	–

Intelligibility is measured by WER achieved by a trained ASR and human perceiver (WER-h). Additional quality measures include STOI, PESQ and MCD. We include results reported in Saleem et al. (2022), Wang et al. (2022), Mira et al. (2022) and Vougioukas et al. (2019) for reference only and acknowledge that performance evaluation was derived from a different dictionary (involving more distinctive words as explained in Sect. 1). Higher values are preferred for metrics marked with (↑); converse is true for (↓)



(a) Plain

(b) Clear

Fig. 3 Confusion matrices of test tokens of **a** plain when trained on plain lip movements and **b** clear when trained on clear lip movement speech styles based on ASR. The numbers shown are percentage responses

8 kHz. The results show that our method achieved better MCD (5.03 at 16 kHz vs 6.51 at 8 kHz) at 16 kHz at the cost of PESQ (1.37 at 16 kHz vs 1.46 at 8 kHz) when compared to the 8 kHz sampling rate.

4.1.2.2 Comparison to the latest method

Table 3 compares the evaluation results from different studies. Top performance (lowest WER) was observed in Saleem et al. (2022), although this study reported results on data where the talkers were already part of training, rendering the task less challenging. The next best was Mira et al. (2022), which was shown to outperform our study which in turn outperformed Akbari et al. (2018). It should be noted that these WER values may not be directly comparable due to use of different dictionaries in these studies. However, comparisons of the additional quality metrics achieved may be informative. For example, comparing our method for 16 kHz with Mira et al. (2022) gave similar perceptual quality (PESQ: 1.37 vs. 1.47), higher intelligibility (STOI: 0.598 vs. 0.523), and lower distortion (MCD: 6.51–6.78 vs. 37.91).fusions made by our ASR are shown in Fig. 3. As seen from the diagonal elements, most tokens are identified correctly. Overall, irrespective of the speech style, "keyed" is being confused with "kid"; "cod" is being confused with "cud". These confusions are expected since the vowels ("keyed"–"kid" and "cod"–"cud") form tense-lax pairs and have been reported in Tang et al. (2015) to have

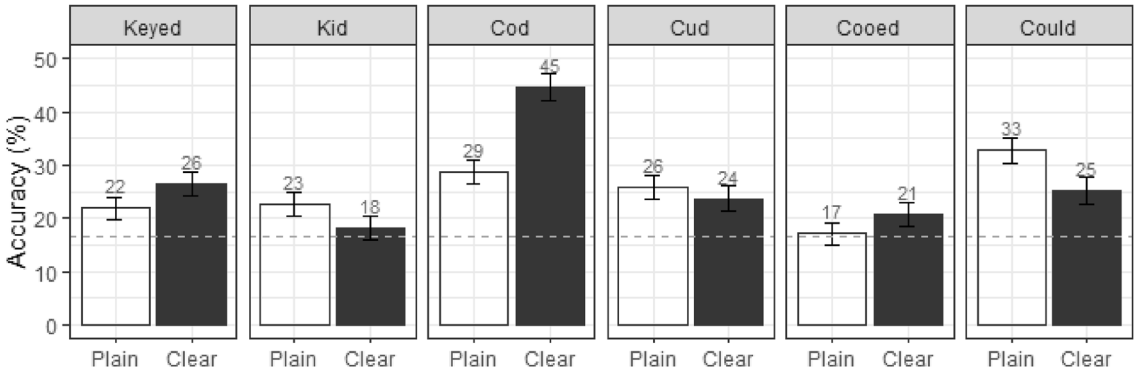


Fig. 4 Identification accuracy (%) for each of the six target words separated by style (plain, clear). The dotted line represents the chance level of 16.7%. The bar whiskers represent SD

similar articulatory movements. Tense-lax confusions are also more frequent for human perceivers than other vowel pairings (Redmon et al., 2020).

When comparing the two confusion matrices, the performance of ASR in the clear speech style has better accuracy than in the plain speech style. In both speech styles, accuracy for the audio synthesis of the three words with the vowels (/i, ɑ, u/) ("keyed", "cod", "cooed") was higher than the other three words with the vowels (/ɪ, ʌ, ʊ/) ("kid", "cud","could"). Interestingly, the former vowels (/i, ɑ, u/) are tense vowels, while the latter three (/ɪ, ʌ, ʊ/) are their lax counterparts. This indicates that the ASR was able to label words with tense vowels with much greater accuracy than those with lax vowels, further suggesting that the deep network achieved a better model of words with tense vowels than lax vowels. Overall, "cooed" and "cod" performed best in both plain speech and clear speech. Further, "could" and "cud" were confused the most in both plain speech and clear speech.

4.2 Experiment 2—evaluation by human perceivers

4.2.1 Methods

Intelligibility of the auto-generated audio from the lip-movements of the video was evaluated by 50 adult native English speakers with self-reported normal hearing and vision recruited via Amazon Mechanical Turk. The experiment was conducted online and was created using a custom version of jsPsych-6.1.0 and put on the JATOS server. In the evaluation, perceivers were asked to listen to and identify the generated audio test samples in a 6-alternative forced choice identification task (i.e., choose one from the six target words). Participants also rated their confidence in their answer on a scale of 1 (not sure) to 5 (very sure).

A total of 860 different generated tokens were included in the experiment. These tokens were divided into 10 sets of 86 tokens (6 words × 2 styles × 6 talkers + 14 remaining tokens), each containing the six target words by six different talkers (3 male, 3 female). Each set was evaluated by 5 different perceivers, and each perceiver was asked to listen to one set. Before the experiment, each perceiver had nine practice trials, which were independently generated and different from the test set of 860 tokens.

4.2.2 Results

The dataset was submitted to sets of generalized linear mixed-effects models using the 'lme4' package in R. Two separate analyses were performed. In the first set of analyses, "Word" and "Style" were included as main fixed effects to examine the intelligibility of each of the synthesized vowels (words) in each speech style. The second set of analyses further examined the effects of vowel "Tensity" and talker "Gender" along with speech "Style", based on the previous articulatory and acoustic findings that interaction of these factors may influence vowel intelligibility (Cutler et al., 2004; Ferguson, 2012; Leung et al., 2016; Tang et al., 2015).

4.2.2.1 Word and style

In the first model, we analyze the overall intelligibility of each synthesized word in each style, as well as the interaction between speech style and the target words. The fixed effects include Word ('keyed', "kid", "cod", "cud", "cooed" and "could") and Style (plain, clear). The dependent variable is word identification accuracy (correct, incorrect). A random effect was added on the intercept term to account for different talkers and perceivers. The linear mixed effect model formula is: Word Identification Accuracy ~ Word*Style + (1|Talker) + (1|Perceiver). Figure 4 displays the comparisons of the identification accuracy in these conditions.

Overall, the average accuracy across the six words is 25.9% (SD = 43.8%), with a range from 36.2% ("cod") to

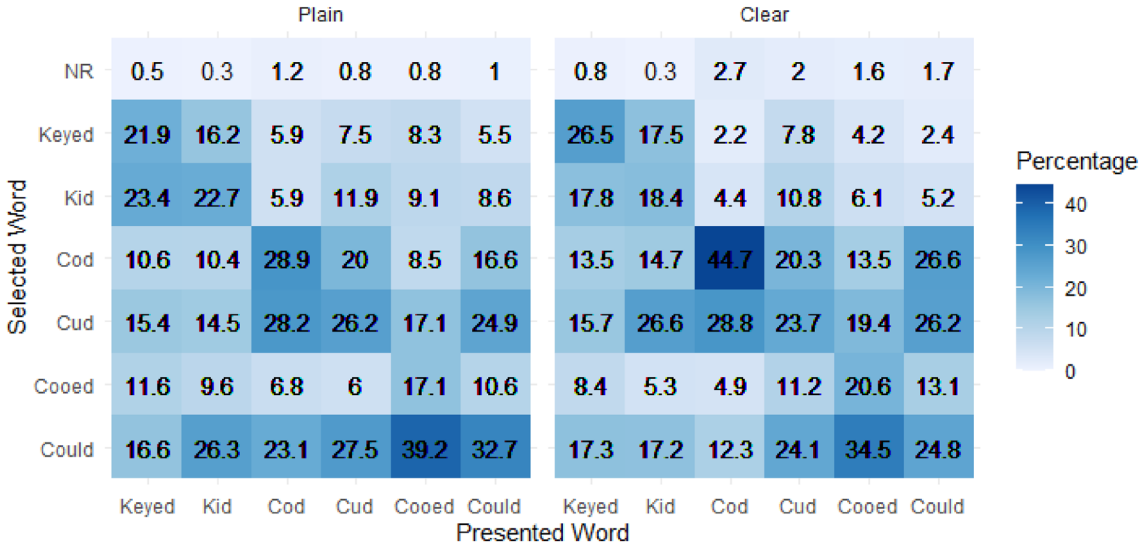


Fig. 5 Confusion matrix of plain and clear target words by human perceivers. The numbers shown are percentage responses. "NR": no response

"18.7% ("cooed"), all significantly above the chance level of 16.7%. The mixed effect analysis shows a significant main effect of Word (χ 2(5) = 31.12, $p < 0.0001$), and a significant interaction between Word and Style (χ 2(5) = 30.09, $p < 0.0001$).

Post-hoc pairwise comparisons among the words were analyzed using the 'emmeans' package in R, with adjusted p-values for multiple comparisons (here and in subsequent sections). The results show that "cod" was significantly more intelligible than the other five words ($p < 0.05$). Apart from "cod", the comparisons reveal significantly greater accuracy for "could" than both "cooed" (odds ratio = 0.558, $p < 0.0001$, $z = -4.41$, CI (0.38, 0.81)) and "kid" (odds ratio = 0.627, $p < 0.01$, $z = -3.57$, CI = (0.43, 0.91)).

Furthermore, post-hoc analyses following the interaction of Word and Style show a significant difference between plain and clear speech for "cod" (odds ratio = 0.49, $p < 0.001$, $z = -4.73$, CI = (0.30, 0.78)), being more intelligible in clear (44.7%) than plain (28.7%) speech. Moreover, in clear speech, the accuracy for "cod" was significantly greater than that for the other five words ($p < 0.05$). In plain speech, "cod" was significantly more accurate than "cooed" (odds ratio = 2.01, $p = 0.002$, $z = 3.96$, CI = (1.16, 3.47)), and "could" was also more accurate than "cooed" (odds ratio = 0.41, $p < 0.0001$, $z = -5.00$, CI = (0.24, 0.72)).

As shown in the confusion matrix in Fig. 5, in most cases, the diagonal elements had the largest value, suggesting each of the target words was selected more frequently than the other word options. In cases where we do have a larger value for an off-diagonal element, that vowel forms a tense-lax pair with the intended vowel, such as "keyed/kid", or "cooed/could". Since tense-lax pairs have similar articulatory movements (e.g., horizontal lip stretching for both "keyed" and "kid") differing primarily in length and degree of articulation (Picheny et al., 1986, Tang et al., 2015), they are easily confusable. These results indicate that the network is able to learn from lip movements of the more contrastive vowels (e.g., "keyed", which involves greater horizontal lip stretching versus "cod", which involves greater vertical lip movement).

In addition to intelligibility accuracy, analyses were also conducted on the confidence rating data, on a scale of 1 (not sure) to 5 (very sure). The mean confidence rating

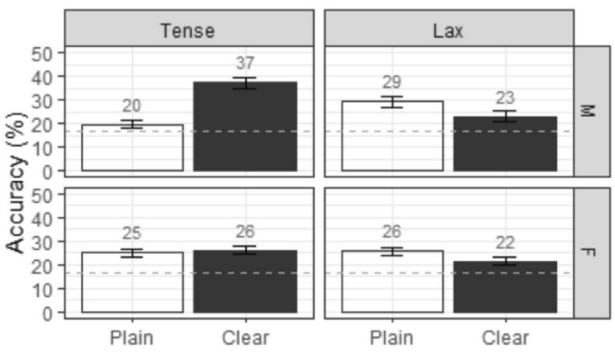


Fig. 6 Identification accuracy (%) for the test tokens by human perceivers as a function of vowel tensity (tense, lax), speech style (plain, clear), and talker gender (male, female). The horizontal dotted line represents chance level

score for correctly identified words across Word and Style was 3.7 (SD = 1.02), which was well above average (3.0). Linear mixed effect modeling (Confidence level ~ Word* Style + (1|Talker) + (1| Perceiver)) showed no significant main effect of Word or Style, nor any interaction between the two. These rating results indicate that perceivers were uniformly confident in their identification of all the synthesized words in both plain and clear speech.

4.2.2.2 Style, tensity and gender

Given the observed confusions between tense and lax vowels, the second set of analyses examined the effects of vowel Tensity (tense, lax) along with speech Style (plain, clear) and talker Gender (male, female) on word intelligibility. Linear mixed-effect analyses was performed with word identification accuracy as the dependent variable. A random effect was added on the intercept term to account for different words and perceivers. The model formula was: Word Identification Accuracy ~ Style * Tensity * Gender + (1|Word) + (1| Perceiver).

As displayed in Fig. 6, modeling results show significant main effects of (1) Style (χ 2(1) = 37.20, p < 0.0001), with greater accuracy for clear speech (27.0%, SD = 44.4%) than plain speech (24.9%, SD = 43.3%); (2) Tensity (χ 2(1) = 4.03, p = 0.04), with greater accuracy for tense vowels (26.6%, SD = 44.2%) than lax vowels (25.0%, 43.3%); and (3) Gender (χ 2(1) = 5.43, p = 0.019), with greater accuracy for male talkers (27.2%, SD = 44.5%) than female talkers (24.8%, SD = 43.2%). In addition, significant interactions between Style and Tensity (χ 2(1) = 31.72, p < 0.0001), Style and Gender (χ 2(1) = 18.52, p < 0.0001), and Tensity and Gender (χ 2(1) = 7.15, p < 0.01) were observed. Further, a significant 3-way interaction was observed between Style, Tensity and Gender (χ 2(1) = 10.78, p < 0.01).

Post-hoc pairwise comparisons revealed two sets of significant differences. Firstly, tense vowels by male talkers show that clear style (37.2%, SD = 48.4%) had significantly greater accuracy than the plain style (19.6%, SD = 39.8%) [odds ratio = 0.40, p < 0.0001, z = − 6.10, CI = (0.27, 0.61)]. Secondly, clear tense vowels produced by male talkers (37.2%, SD = 48.4%) were identified with greater accuracy than those produced by female talkers (26.2%, SD = 44.0%) [odds ratio = 1.69, p = 0.002, z = 3.77, CI = (1.14, 2.49)).

In summary, the human intelligibility results reveal that the accuracy of all the synthesized words was above chance. In terms of individual words, "cod" was found to be the most intelligible, followed by "could", whereas "cooed" was the least intelligible. Notably, word confusions were largely seen between tense and lax vowel pairs. The identification accuracy of synthesized words was also found to be affected by speech style, talker gender, and vowel tensity, with clear speech being more accurately identified than plain speech for tense-vowel words by male talkers.

5 Discussion

5.1 Video-to-audio vowel synthesis and intelligibility

The above-chance performance for all the synthesized words, indicates that an audio speech signal can be recreated based on articulatory movements extracted from a talker's face. Such cross-modal synthesis demonstrates a direct link between visual articulatory and acoustic cues, in that facial movements characterizing different vowels (e.g., lip spreading, jaw lowering, lip-rounding) (Kim & Davis, 2014; Tang et al., 2015; Tasko & Greilick, 2010) can be translated into formant patterns to produce intelligible audio speech. This is in line with the previous claim of positive correlations among articulation, acoustics, and intelligibility measures of speech sounds (Gagné et al., 2002; Kim & Davis, 2014; Tasko & Greilick, 2010).

Indeed, the current intelligibility results of the video-to-audio synthetic words are consistent with the patterns from the perception of naturally produced audio or visual input, including patterns due to the effects of speech style, talker gender, as well as tensity of articulatory gestures (Heald & Nusbaum, 2014; Kim & Davis, 2014; Redmon et al., 2020; Tasko & Greilick, 2010; Traunmüller & Öhrström, 2007).

In particular, the vowel confusion patterns show that most of the confusions are between tense and lax vowel pairs, presumably due to their articulatory and acoustic similarities, as has also been reported in previous studies in both auditory and visual domains (e.g.,Cutler et al., 2004; Redmon et al., 2020; Tang et al., 2015). This further suggests that the visual features distinguishing tense and lax vowels, including extent of the articulation and duration, may not robustly contribute to distinctive audio cues for tense and lax vowels, unless these visual features are enhanced in clear speech.

Vowel tensity has been found to interact with clear speech characteristics, in that the synthetic tense vowels but not lax vowels exhibit a clear speech benefit in intelligibility. In previous work with natural stimuli, visual perception of tense vowels exhibits a clear speech advantage while that of lax vowels demonstrates a disadvantage; although in auditory perception, such a clear speech benefit is found for both tense and lax vowels (Redmon et al., 2020). This is presumably because clear speech modifications, which involve more extreme articulatory gestures, are compatible with the inherent features of tense vowels and thus benefit intelligibility, but are in conflict with lax vowel features and hinder intelligibility (Hillenbrand

et al., 1995; Lam et al., 2012; Roesler, 2013; Smiljanic & Bradlow, 2009; Tang et al., 2015). As such, the current results demonstrate direct perceptual consequences of cross-modal synthesis, in that auditory perceptual patterns reflect how the synthesis network was built. If the network were based on acoustic features, then clear speech would have been equally beneficial for tense and lax vowels. Since the current audio-based perception shows a clear speech benefit with synthesized tense vowels only, this suggests that the network was able to learn from tense vowel articulatory features across styles to produce the corresponding audio, but not when clear-speech modifications confounded visual cues to lax vowels.

Moreover, vowel tensity and speech style also interact with talker gender, in that a clear speech advantage was only observed for tense vowels produced by male talkers. Previous articulatory research on these vowels has shown that male compared to female talkers have larger clear-plain distinctions in visual articulatory movements (Tang et al., 2015), whereas male and female talkers do not differ in the acoustic patterns of clear speech modifications (Leung et al., 2016). Thus the current results demonstrate that perception of the synthetic vowels reflects how the vowels were synthesized. As male talkers' clear speech articulation involves greater articulatory changes compared to female talkers', the network was able to better extract such visible articulatory variations from male than female talkers and associate those features with the corresponding audio, leading to greater intelligibility.

These results have significant implications for cross-modal synthesis, in that such synthesis should take into account, and take advantage of, specific characteristics in different (articulatory and acoustic) domains and across talkers, which may consequently maximize intelligibility benefits.

5.2 ASR and human evaluation

ASR systems have been claimed to achieve human-like performance in human speech classification (Xiong et al. 2018) and are thus widely used in speech synthesis evaluations (Akbari et al., 2018; Mira et al., 2022; Saleem et al., 2022; Vougioukas et al., 2019).

The current results of WER analyses show that ASR even outperformed human listeners in classifying the target words. The performance of the ASR system depends heavily on the dataset it is trained on. The current audio synthesis network and ASR both use formant frequencies and frequency-based features in loss function to classify the words. Thus, the basis for ASR classification is well mapped with the synthesis network. ASR can learn to recognize subtle differences between audio features to classify an audio signal in a multi-class classification problem. This may lead to improved performance, as revealed by the current ASR results, compared to human evaluation.

However, current state-of-the-art ASRs also struggle with large variations in speech due to speaker characteristics (e.g., gender, accent) and linguistic factors (e.g., speech context, word frequency) (Feng et al., 2021). In contrast to ASR, in human speech perception and word recognition, perceivers not only draw on the input signal but also rely on their prior experience which may involve a variety of speech and non-speech cues. This difference may explain the discrepancies in the classification accuracy of individual words between the ASR and human results. For example, for ASR, the classification of "cooed" was highly accurate while that of "could" was poor. In contrast, for human perceivers, "could" was among the more intelligible words while "cooed" was the least intelligible. This could be because "cooed" is a less common word for human perceivers that "could". The confusion patterns further reveal that "cooed" was most frequently perceived by human perceivers as "could". On the other hand, in our balanced dataset, word frequency did not affect ASR, as "could" was most frequently confused with "cod".

Taken together, the present study suggests that ASR, as an effective tool for evaluating the performance of the audio synthesis network, can be used in preliminary evaluation of the network's performance and hyperparameters tuning. However, human evaluation should also be included for word classification that takes into account human experiential factors. Thus, the current findings suggest that ASR and human intelligibility can work in a complementary manner in developing accurate and naturalistic cross-modal speech synthesis networks.

6 Concluding remarks and future directions

This research is the first attempt to conduct cross-modal video-to-audio speech synthesis involving words that are minimally contrastive in vowels. Our approach is unique compared to previous methods in a number of ways. To start, previous synthesis work primarily focused on sentence-level information (Akbari et al., 2018; Vougioukas et al., 2019), thus providing redundant cues for word distinction. The current vowel-based synthesis with minimally contrastive words makes the lip-reading task exceptionally challenging, since word recognition is only based on cues in a single speech segment: the vowel. Further, this segment-based method allows us to extend the previous research by delving further into variations of speech due to vowel tensity, speech style and talker gender differences in lip reading, showing that these factors can be incorporated by the network to synthesize audio, which were not considered in previous methods. Additionally, using standard landmark points instead of directly using a talker's

face (as previously adopted) makes feature extraction (facial) texture-independent and speaker-independent. Lastly, the network performance was evaluated in a real-life scenario where test talkers were not part of the training set, unlike previous work (e.g., Akbari et al., 2018). All in all, the factors employed by the current approach suggest that our synthesis network is highly generalizable across contextual and talker variations and in natural communicative settings.

Results from this study point to promising directions for future work on cross-modal speech synthesis that integrates deep learning and linguistic approaches. First, while the current research focused on the trajectories of articulatory movements, adding temporal information to the movements in training will help improve network performance, as duration typically differs between tense and lax vowels and between plain and clear speech (Leung et al., 2016; Picheny et al., 1986; Smiljanic & Bradlow, 2009). Moreover, to improve the word identification accuracy, a profile face could be included, as certain articulatory features such as lip protrusion are better identified on the profile face (Tang et al., 2015). It is conceivable that having both frontal and profile face information will improve performance. Further, in this study, we chose to start with a set of vowels with distinctive visible articulatory movements. Future work can extend the current framework to systematically synthesize consonants as well as vowels, with the goal of achieving a comprehensive segment-based synthesis network independent of speech contexts and styles. Another potential future direction for research is to develop and refine protocols for comparing different methods using common corpora for testing.

Acknowledgements This research has been supported by the Big Data Next Big Question Fund at Simon Fraser University (SFU) and a grant from the Social Sciences and Humanities Research Council of Canada (SSHRC Insight Grant 435-2019-1065). We thank Shubam Sachdeva, Jetic Gu, Keith Leung, Lisa Tang, and members of the Language and Brain Lab at SFU for their assistance.

Data availability The datasets generated during and analysed during the current study are available from the corresponding author on reasonable request. The code used during the current study is available in the repository: https://github.com/srbhgarg/VowSynth.git.

References

Akbari, H., Himani A., Cao, L. & Mesgarani, N. (2018). Lip2Audspec: Speech reconstruction from silent lip movements video. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (pp. 2516–2520). https://doi.org/10.1109/icassp.2018.8461856.

Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2018). Intelligible speech synthesis from neural decoding of spoken sentences. *BioRxiv*. https://doi.org/10.1101/481267

Assael, Y. M., Shillingford, B., Whiteson, S., & De Freitas, N. (2016). Lipnet: End-to-end sentence-level lipreading. *arXiv preprint* arXiv:1611.01599.

Bernstein, L. E., Auer, E. T., Jr., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication, 44*(1–4), 5–18.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9), 341–345.

Bond, Z. S., & Moore, T. J. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication, 14*, 325–337. https://doi.org/10.1016/0167-6393(94)90026-4

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication, 20*(3–4), 255–272.

Burris, C., Vorperian, H. K., Fourakis, M., Kent, R. D., & Bolt, D. M. (2014). Quantitative and descriptive comparison of four acoustic analysis systems: Vowel measurements. *Journal of Speech, Language, and Hearing Research, 57*(1), 26–45.

Chen, L., Su, H., & Ji, Q. (2019). Deep structured prediction for facial landmark detection. *Advances in Neural Information Processing Systems, 32*, 158.

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America, 120*(5), 2421–2424.

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America, 116*(6), 3668–3678.

Dubnov, S. (2004). Generalization of spectral flatness measure for non-gaussian linear processes. *IEEE Signal Processing Letters, 11*(8), 698–701.

Ephrat, A., & Peleg, S. (2017). Vid2speech: speech reconstruction from silent video. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5095–5099). IEEE.

Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021). Quantifying bias in automatic speech recognition. *arXiv preprint* arXiv:2103.15122.

Ferguson, S. H. (2012). Talker differences in clear and conversational speech: Vowel intelligibility for older adults with hearing loss. *Journal of Speech Language and Hearing Research, 55*(3), 779–790.

Ferguson, S. H., & Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America, 112*, 259–271.

Ferguson, S. H., & Kewley-Port, D. (2007). Talker differences in clear and conversational speech: Acoustic characteristics of vowels. *Journal of Speech Language and Hearing Research, 50*, 1241–1255.

Ferguson, S. H., & Quené, H. (2014). Acoustic correlates of vowel intelligibility in clear and conversational speech for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America, 135*(6), 3570–3584.

Freitas, J., Teixeira, A., Dias, M. S., & Silva, S. (2017). *An introduction to silent speech interfaces*. Springer.

Gagné, J. P., Rochette, A. J., & Charest, M. (2002). Auditory, visual and audiovisual clear speech. *Speech Communication, 37*(3–4), 213–230.

Garg, S., Tang, L., Hamarneh, G., Jongman, A., Sereno, J. A., & Wang, Y. (2019). Computer-vision analysis shows different facial movements for the production of different Mandarin tones. *The Journal of the Acoustical Society of America, 144*(3), 1720–1720.

Gonzalez-Lopez, J. A., Gomez-Alanis, A., Doñas, J. M. M., Pérez-Córdoba, J. L., & Gomez, A. M. (2020). Silent speech interfaces for speech restoration: A review. *IEEE Access, 8*, 177995–178021.

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 32*(2), 236–243.

Harte, C., Sandler, M., & Gasser, M. (2006). Detecting harmonic change in musical audio. In *Proceedings of the 1st ACM workshop on audio and music computing multimedia* (pp. 21–26).

Heald, S., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience, 8*, 35.
Herff, C., Heger, D., De Pesters, A., Telaar, D., Brunner, P., Schalk, G., & Schultz, T. (2015). Brain-to-text: Decoding spoken phrases from phone representations in the brain. *Frontiers in Neuroscience, 9*, 217.
Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97*(5), 3099–3111.
Hueber, T., Benaroya, E. L., Chollet, G., Denby, B., Dreyfus, G., & Stone, M. (2010). Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication, 52*(4), 288–300.
Jongman, A., Wang, Y., & Kim, B. H. (2003). Contributions of semantic and facial information to perception of nonsibilant fricatives. *Journal of Speech Language and Hearing Research, 46*, 1367–1377.
Kawase, T., Hori, Y., Ogawa, T., Sakamoto, S., Suzuki, Y., & Katori, Y. (2015). Importance of Visual Cues in Hearing Restoration by Auditory Prosthesis. In *Interface Oral Health Science 2014* (pp. 119–127). Springer
Kim, J., & Davis, C. (2014). Comparing the consistency and distinctiveness of speech produced in quiet and in noise. *Computer Speech & Language, 28*(2), 598–606.
King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research, 10*, 1755–1758.
Laitinen, M. V., Disch, S., & Pulkki, V. (2013). Sensitivity of human hearing to changes in phase spectrum. *Journal of the Audio Engineering Society, 61*(11), 860–877.
Lam, Jennifer, Tjaden, Kris, & Wilding, Greg (2012). Acoustics of clear speech: Effect of instruction. *Journal of Speech Language and Hearing Research 55*(6), 1807–1821. https://doi.org/10.1044/1092-4388(2012/11-0154)
Le Cornu, T., & Milner, B. (2015). Reconstructing intelligible audio speech from visual speech features. In *Interspeech* (pp. 3355–3359).
Leung, K. K., Redmon, C., Wang, Y., Jongman, A., & Sereno, J. (2016). Cross-linguistic perception of clearly spoken English tense and lax vowels based on auditory, visual, and auditory-visual information. *The Journal of the Acoustical Society of America, 140*(4), 3335–3335.
Lu, Y., & Cooke, M. (2008). Speech production modifications produced by competing talkers, babble, and stationary noise. *The Journal of the Acoustical Society of America, 124*(5), 3261–3275.
Maniwa, K., Jongman, A., & Wade, T. (2008). Perception of clear fricatives by normal-hearing and simulated hearing-impaired listeners. *The Journal of the Acoustical Society of America, 123*, 1114–1125.
Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America, 125*(6), 3962–3973.
Mira, R., Vougioukas, K., Ma, P., Petridis, S., Schuller, B. W., & Pantic, M. (2022). End-to-end video-to-speech synthesis using generative adversarial networks. In *IEEE transactions on cybernetics*. arXiv:2104.13332 [cs.LG]
Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science, 15*(2), 133–137.
Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America, 24*(2), 175–184.
Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing II. *Journal of Speech Language and Hearing Research 29*(4), 434–446. https://doi.org/10.1044/jshr.2904.434
Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020). Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13796–13805).
Redmon, C., Leung, K., Wang, Y., McMurray, B., Jongman, A., & Sereno, J. A. (2020). Cross-linguistic perception of clearly spoken English tense and lax vowels based on auditory, visual, and auditory-visual information. *Journal of Phonetics, 81*, 100980.
Roesler, L. (2013). *Acoustic characteristics of tense and lax vowels across sentence position in clear speech*. Unpublished Master's thesis, University of Wisconsin-Milwaukee
Saleem, N., Gao, J., Irfan, M., Verdu, E., & Fuente, J. P. (2022). E2E–V2SResNet: Deep residual convolutional neural networks for end-to-end video driven speech synthesis. *Image and Vision Computing, 119*, 104389.
Savitzky, A., & Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*(8), 1627–1639.
Schultz, T., & Wand, M. (2010). Modeling coarticulation in EMG-based continuous speech recognition. *Speech Communication, 52*(4), 341–353.
Smiljanić, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistics Compass, 3*(1), 236–264.
Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212–215.
Tang, L. Y., Hannah, B., Jongman, A., Sereno, J., Wang, Y., & Hamarneh, G. (2015). Examining visible articulatory features in clear and plain speech. *Speech Communication, 75*, 1–13.
Tasko, S. M., & Greilick, K. (2010). Acoustic and articulatory features of diphthong production: A speech clarity study. *Journal of Speech Language and Hearing Research, 53*, 84–99.
Traunmüller, H., & Öhrström, N. (2007). Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics, 35*(2), 244–258.
van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. In *Proceeding of 9th ISCA workshop on speech synthesis workshop (SSW 9),* 125
Vougioukas, K., Ma, P., Petridis, S., & Pantic, M. (2019). Video-driven speech reconstruction using generative adversarial networks. *arXiv preprint* arXiv:1906.06301.
Wang, Disong, Yang, Shan, Su, Dan, Liu, Xunying, Yu, Dong & Meng, Helen. (2022). VCVTS: Multi-speaker video-to-speech synthesis via cross-modal knowledge transfer from voice conversion.
Watson, C. I., & Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *The Journal of the Acoustical Society of America, 106*(1), 458–468.
Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018). The Microsoft 2017 conversational speech recognition system. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5934–5938). IEEE.
Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26*(1–2), 23–43.
Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters, 23*(10), 1499–1503.