Reading help for the visually impaired: the case of talking ads

Stefan Maxim
The Harker School
San Jose, Ca
stefan.teodor.maxim@gmail.com

Abstract

Most existing grocery ads for visually impaired people are processed and recited manually, which can be very time-consuming. Furthermore, there is a lack of appropriate tools available for optimizing and personalizing the audio. This paper proposes an automated method for generating audio ads for visually impaired people by transforming the traditional paper or online ads into an enhanced audio version that selectively speeds up audio portions based on the pitch of the words, all while organizing them based on categories. The generated audio files take advantage of the fact that visually impaired people are more capable of hearing high-pitched words and can process audio information at much higher speeds (2X to 3x) than most sighted people. By re-training a pre-trained model for image detection alongside a classical code base for text-to-audio and audio augmentation, we were able to vertically integrate the process of reading ads in a more compact and intuitive way. This has resulted in a method that is not only more accessible for those with visual impairments but still remains portable and light enough to be utilized on mobile devices.

1 Introduction

For most people, AI and technology have greatly increased the rate of consumption of various visual media types like articles, newspapers, and journals. However, the same cannot be said about visually impaired people, as they are forced to find a different approach to interpreting media. Currently, approximately 304.1 million Rupert R A Bourne [2020] people have moderate to severe blindness while around 49.1 million are completely blind, a stark 42.8 % increase from the 34.4 million in 1990. While medical research is helping reduce the number of preventable cases, challenges like access to information and mobility remain a problem. Specifically, there is a gap of 55 words per minute (wpm) between the average reading rate (238wpm) and speaking rate (183wpm) per Rupert R A Bourne [2020]. This results in a 23 % increase in time spent by a person who consumes audio-only information. In response, this paper proposes a method to automate the process by which grocery adverts (example ad in Figure 1.a) are converted to audio form while decreasing the time to consume.

Figure 1.a illustrates the top-level diagram of the conversion application. It begins by processing the image and then uses ML detection to denote the picture, text, and price. It then classically sorts the ads, indexing at text level to create a searchable database organized by pre-established criteria such as the type of object or price. A selective speed-up of the low-pitch words is performed as they are split into phonemes. This is followed by the text-to-audio conversion using variable speed based on the user feedback and neurological research of the visually impaired. Figure 1.b shows the simplified diagram of the YOLOv8n (nano) single-pass real-time object detection region-based convolutional neural network that uses anchor-free detection, decoupled head and modified loss function for processing speed-up.

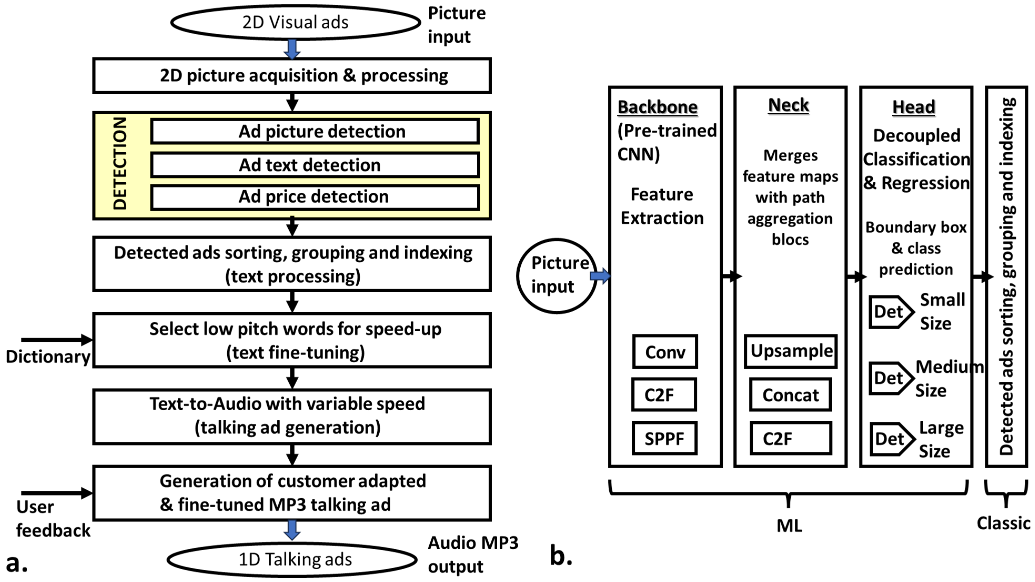
2D Visual ads
Picture input
2D picture acquisition & processing
DETECTION
Ad picture detection
Ad text detection
Ad price detection
Detected ads sorting, grouping and indexing (text processing)
Dictionary
Select low pitch words for speed-up (text fine-tuning)
Text-to-Audio with variable speed (talking ad generation)
User feedback
Generation of customer adapted & fine-tuned MP3 talking ad
1D Talking ads
Audio MP3 output
a.
Picture input
Backbone
(Pre-trained CNN)
Feature Extraction
Conv
C2F
SPPF
Neck
Merges feature maps with path aggregation blocs
Upsample
Concat
C2F
Head
Decoupled Classification & Regression
Boundary box & class prediction
Det
Small Size
Det
Medium Size
Det
Large Size
Detected ads sorting, grouping and indexing
ML
Classic
b.

Figure 1: a. Bloc diagram of the proposed talking ads generation and fine-tuning using both ML and classic processing; b. YOLOv8n real-time single-pass object-detection R-CNN simplified diagram

2 Related work

The most common method of converting the ads to audio form is via a human manually reciting them. Alternatively, one can use existing LLMs like ChatGPT, but it is very high-power and unwieldy for those with blindness. Furthermore, the LLMs output will still need to be recited either manually or automatically at the established 183 wpm, making the total time much slower than the 238 wpm Brysbaert [2019] of reading. Furthermore, when prompted to read an ad page, ChatGPT produced an ordered list of all the words from a page. However, when asked to make its results more useful for those with blindness, the result just becomes more verbose while keeping its sequentiality. This is the same way that ads are often manually red, which is often detrimental to the listeners as people usually skim until seeing parts they are interested in, something that this archaic method makes impossible.

Several papers Kawamura and Rekimoto [2023] propose varying audio speed to increase listening efficiency, however, these solutions are focused on sighted individuals. Many studies like Fields [2010] highlight the fact that blind people can understand speech at speeds 2-3 times faster than sighted individuals. Other studies like Wan et al. [2010], Occelli et al. [2016], and Gougoux et al. [2004] show that congenitally blind people perform better than sighted people on a range of verbal tasks like pitch discrimination (spectral and temporal ability to distinguish pitch changes Bertonati et al. [2021]), auditory memory (short term memory recall of words), and verbal fluency (both semantic and letter fluency). Instead of only focusing on these traits as neurological insights, I decided to use them to tailor the audio produced in my visual-to-audio program to be more efficient and allow greater listening speeds for people with visual impairments.

3 Methodology

Our solution has three parts: automatic verbal ad generation, variable speech adaptation for low-pitch words, and improvements in text-to-audio translation.

3.1 Building the pipeline

For detecting ads, we used the pre-trained model yolov8n because of performance, size (32m parameters), and ease of use Ultralytics. By looking through hundreds of ads, we noticed that most have the same format: a rectangular area with an image, a description, and a price. For our data, we used a combination of “real” images created by manually labeling ads as well as synthetic images (item pictures are sourced from Fruits360 dataset Molnár and generated text). We then built three training datasets: 200 synthetic, 200 real, and a combination of 200 real and 900 synthetic images, and one validation dataset with 50 real images. We are interested in 3 key metrics: the F1 score per class, the mean average precision (both for 0.5 and 0.95 IOU), and the amount of data needed to achieve a good accuracy. Our hypothesis is that our dataset is sufficient to achieve a >0.9 mAP50 accuracy. After the text and the price are detected, we are using Easy OCR Jaided AI to convert them to text and gTTF Google to convert the text into speech.

3.2 Accelerating low pitch words

Since visually impaired individuals are able to discern high-pitch sounds, we hypothesize that selectively increasing the speed of low-pitch words should maintain the intelligibility of a text. For each ad text, we decompose it into words and then further down into phonemes. For the words with low-pitch phonemes, we increased the speed of pronunciation. We created a control sample with all words at a constant speed (1) and several variable speed ads with speed values of 1.3X, 1.5X, 1.7X, 2.0X, and 2.7X respectively. We tested the resulting audio files on 33 sighted individuals. Each participant was given ads at both constant and variable speed and was asked to evaluate the audio on a 1-10 scale where 1 meant unintelligible and 10 meant clear.

3.3 2D to 1D transformation

We were then tasked with efficiently transforming the image (2D) into audio (1D). Our goal was to answer “what information can be eliminated from the 2D space when moving it to a 1D space?”. The 33 sighted individuals who participated in our study were asked to describe how they read the weekly ads, how long it takes, and what they remember from it.

4 Results

We will start with the results from Section 3.1.Our experiments (source code, data and more results: Maxim, S) returned 6 models, each with a different combination of dataset size/composition and epochs, and obtained the mAP50 scores and mAP50-95 depicted in table 1

mAP scores for each model			
Model Name	Notes	mAP50	mAP50-95
Model 1	200 synthetic images for 10 epochs	0.82705	0.46586
Model 2	200 real images for 10 epochs	0.97442	0.75793
Model 3	1100 hybrid images for 10 epochs	0.**98654**	**0.76533**
Model 4	200 synthetic images for 50 epochs	0.91534	0.54357
Model 5	200 real images for 50 epochs	0.94266	0.7137
Model 6	1100 hybrid images for 50 epochs	0.98649	0.73275

Table 1: Models with mAP50 and mAP50-90 metrics

Model 1, trained on synthetic data obtained a mAP50-95 score of approximately 0.46, which was less than that of the purely manual data Model 2, which obtained a value of 0.76. In Model 4, by increasing the epochs from Model 1, the corresponding mAP score increased to 0.54, an increase of around 0.077. Interestingly, in Model 5, the same increase in epochs from its predecessor Model 2 actually lowered the mAP score to 0.71, a decrease of around 0.044. This pattern held true with the composite dataset models(namely models 3 and 6), albeit to a less severe degree as the mAP score only decreased by 0.032, from around 0.77 to 0.73.

Lastly, we depicted the F1 curve for the Best Model (Model 3) as well as that of Model 2, the second best, side by side in Figure 2. Qualitatively, we can see that the mAP, as well as the overall F1 score, seems to be tighter as well as higher for our best model.

Analyzing the Section 3.2 data, it is observed that people can easily understand ads at speeds up to 1.5x with an average understanding score of 7.636 out of 10, yet this number drops to 4.182 for any higher speeds.

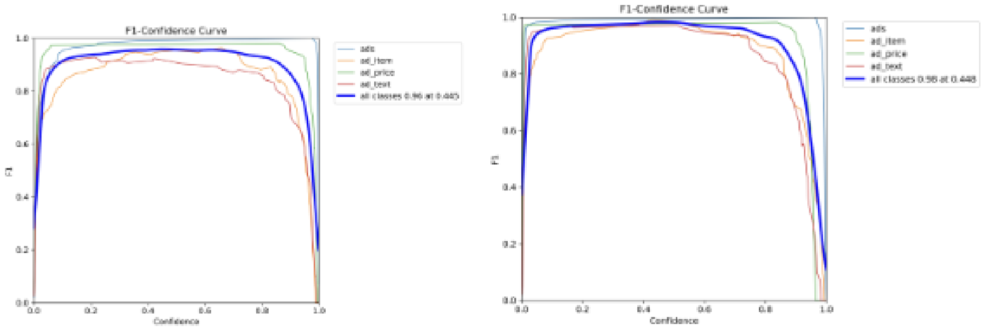

Figure 2: F1 curve for best and model trained with only real data

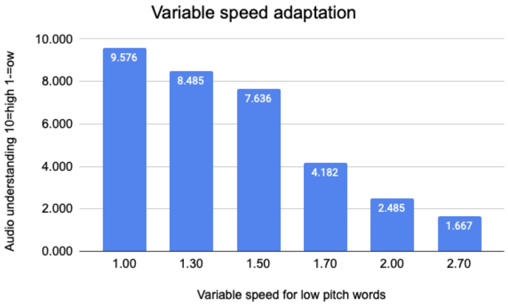

Figure 3: Understanding of an audio ad that has words at variable speed

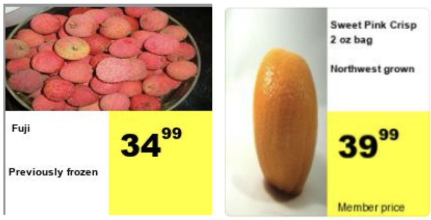

Figure 4: Sample synthetic data

Our results in Section 3.3 found that people spend an average of 20 seconds per ad due to skimming the visuals and focusing on the category of interesting items. They then remembered 2-3 items per page.

5 Discussion

The model we trained that took an image as input and detects the ad image, text, and price demonstrates the feasibility of our automated ad detection. It showed that the most effective way of training the model while minimizing labor costs would be a combination of real and synthetic data, thereby confirming the efficacy of our synthesized data (highlighted in Man and Chahl [2022] as well). We were also able to validate our hypothesis that speeding up low-pitch words in a sentence does not lead to information loss. The results show that speeds up to 1.5x do not significantly change the understanding of the content, which for visually impaired individuals who re-play the ad at even higher speeds, means the efficiency in ad listening is even greater. Lastly, we highlighted the need to process the text based on grouping rather than the current sequential processing. Our proposed method of grouping the items is in line with how ads are actually visualized.

6 Conclusion and future work

We have studied the detection accuracy as a function of the dataset size, using both real grocery ads and synthetically generated ones. To ensure detection accuracy above the 95 threshold level, the training data needs to contain at least 1k synthetic data samples. When compared with the manually generated audio ads for visually impaired people, the proposed adaptive speed and selective emphasis of low-pitch words have resulted in at least 1.5x increase in speed for certain words. Using the automated ads grouping and index listing can result in 5 to 10x speed-up of the ad listening time, by jumping directly to the sections of interest. Due to the nature of grocery ads that are rectangular-based, these results can be generalized internationally. Additionally, an extension of this work can focus on improving further the 2D(image) to 1D(audio) translation at the full page level.

References

G. Bertonati, M. B. Amadeo, C. Campus, and M. Gori. Auditory speed processing in sighted and blind individuals. *PLOS ONE*, 16(9):e0257676, 2021. doi: 10.1371/journal.pone.0257676. URL `https://doi.org/10.1371/journal.pone.0257676`.

Marc Brysbaert. How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of Memory and Language*, 109:104047, 2019. ISSN 0749-596X.

R. Douglas Fields. Why can some blind people process speech far faster than sighted persons? *Scientific American*, Dec 13 2010.

Google. GTTS Documentation. `https://gtts.readthedocs.io/en/latest/`.

Frédérique Gougoux, Franco Lepore, Maryse Lassonde, Patrice Voss, Robert J. Zatorre, and Pascal Belin. Neuropsychology: pitch discrimination in the early blind. *Nature*, 430(6997):309, Jul 15 2004. doi: 10.1038/430309a.

Jaided AI. EasyOCR. `https://github.com/JaidedAI/EasyOCR`.

Kazuki Kawamura and Jun Rekimoto. Aix speed: Playback speed optimization using listening comprehension of speech recognition models. In *Proceedings of the Augmented Humans International Conference 2023*, page 200–208, 2023.

K. Man and J. Chahl. A review of synthetic image data and its use in computer vision. *J Imaging*, 8 (11):310, Nov 21 2022. doi: 10.3390/jimaging8110310.

Maxim, S. Code and Documentation. `https://github.com/StefanMaxim/AdsProject`.

B. Molnár. Fruits 360 dataset. `https://www.kaggle.com/datasets/moltean/fruits`.

Valeria Occelli, Simon Lacey, Careese Stephens, and Krish Sathian. Superior verbal abilities in congenital blindness. *Society for Imaging Science and Technology*, 2016. doi: 10.2352/ISSN.2470-1173.2016.16HVEI-094. DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-094.

Seth Flaxman Paul Briant Michele Bottone Theo Vos Kovin Naidoo Tasanee Braithwaite Maria Cicinelli Jost Jonas Hans Limburg Serge Resnikoff Alex Silvester Vinay Nangia Hugh R Taylor Rupert R A Bourne, Jaimie Adelson. Global prevalence of blindness and distance and near vision impairment in 2020: progress towards the vision 2020 targets and what the future holds. In *Invest. Ophthalmol. Vis. Sci.*, page 61(7):2317, 2020.

Ultralytics. Ultralytics Documentation. `https://docs.ultralytics.com`.

Catherine Y. Wan, Amanda G. Wood, David C. Reutens, and Sarah J. Wilson. Early but not late-blindness leads to enhanced auditory perception. *Neuropsychologia*, 48(1):344–348, 2010. ISSN 0028-3932.