Exploring Automated Voice Casting for Content Localization Using Deep Learning

By Aansh Malik and Ha Nguyen

Abstract

Casting voice-actors to dub source language content into a target language—known as voice casting—consists largely of a manual workflow that could benefit immensely from increased levels of automation. Recent advancements in deep learning architectures for sequential data processing are providing the needed impetus to the realization of various AI-enabled audio-processing workflows. Specifically, applications such as speaker verification and speech synthesis have been gaining immense traction due to the advent and maturity of recurrent neural networks. We explore the viability of leveraging advancements in deep learning for text-independent speaker verification (TI-SV) for use in computer-aided voice casting. To this end, we propose and develop an automated voice-casting tool that uses similarity scores generated from neural network embeddings—from a robust autoencoder model trained for the task of TI-SV—to rank voiceover artists across different languages in voice-casting process. To evaluate the dexterity of the proposed approach, we conduct a subjective study emulating a simplified voice-casting process on actual voice-testing kits (dubbing auditions) from our content. We also use casting decisions from casting experts to further evaluate the tool as well as the subjectivity involved in the voice-casting process. We achieve promising results for the automated tool and prove that it could be a viable approach to automating the voice-casting process and warrants further exploration.

Keywords

Artificial intelligence, audio processing, automation, content localization, deep learning, dubbing, machine learning, neural networks, voice casting, voice similarity

Introduction

Content localization is essential for media companies to make their multimedia content including films, games, and television shows, available to global audiences. The audio in the original source language is often translated and replaced with the target language, in a process referred to as dubbing, to provide the highest level of immersion for foreign audiences. Voice casting is the process of selecting voice-actors for dubbing in target languages, which is usually a manual process performed by human experts. The experts have access to audio samples—typically under 60 sec in duration—from original source language characters as well as audio samples from multiple foreign target language voiceover artists. They then evaluate the target language voiceover artists and subjectively cast the most appropriate matches to the original cast. It is important that the voices of casted target language actors share a high acoustic resemblance to those of the original actors in the source language content. Acoustic resemblance is not only important to effectively translate the performance of the original characters, but also to alleviate the audio–visual dissonance audiences experience when watching the original cast on screen but hearing someone else's voice.

Voice casting is the process of selecting voice-actors for dubbing in target languages, which is usually a manual process performed by human experts. The experts have access to audio samples—typically under 60 sec in duration—from original source language characters as well as audio samples from multiple foreign target language voiceover artists. They then evaluate the target language voiceover artists and subjectively cast the most appropriate matches to the original cast. It is important that the voices of casted target language actors share a high acoustic resemblance to those of the original actors in the source language content. Acoustic resemblance is not only important to effectively translate the performance of the original characters, but also to alleviate the audio–visual dissonance audiences experience when watching the original cast on screen but hearing someone else's voice.

If we were able to quantify perceptual voice similarity between two voices in different languages, we would be able to guide the voice-casting process more objectively as

Digital Object Identifier 10.5594/JMI.2021.3057695
Date of publication: 6 April 2021

1545-0279/21©2021SMPTE

well as open avenues for automated voice-casting tools. Perceptual voice similarity for the dubbing process has been studied in Refs. 1 and 2. The authors compare two different techniques of using similarity scores derived from speaker-recognition techniques[3,4] as well as describing a voice using paralinguistic categories speaker states and traits (e.g., age/gender/voice quality and emotions). In Ref. 5 the authors propose an I-vector probabilistic linear discriminant analysis technique to estimate dubbing proximity both in source and target language. Recent advancements in deep neural network embeddings[6–8] and end-to-end speaker-recognition models perform significantly better than the older I-vector approaches for speaker recognition. In fact, the text-independent speaker-verification approach from Ref. 6 has been shown to generalize well to other applications such as voice synthesis.[9]

The relationship between similarity scores from such techniques and human perceptual similarity between two voices is still an open area of research. Moreover, the viability of using similarity scores across different languages (source–target language combinations) for the specific use case of voice casting still lacks concrete evidence. Using similarity scores from speaker-recognition techniques offers an attractive proposition for measuring similarity between voices in the acoustic space for voice casting. There is a general lack of data sets that can be used to train models specific to the voice-casting process. The dubbing data sets that exist contain only final dubbing samples and lack data from voice-casting auditions (known as voice testing kits (VTKs) for the source language and dubbing tests for the target language), which are critical to the voice-casting problem. Speaker recognition, on the other hand, is a heavily researched problem with ample data sets available. It would, therefore, be ideal if similarity scores from state-of-the-art text-independent speaker-recognition models could be directly utilized in the domain of voice casting.

In this study, using our automated voice-casting tool, we attempt to evaluate the use of similarity scores between neural network embeddings from voices in source and target languages for voice casting. We also conduct a subjective experiment to test the performance of the automated tool against human testers. The subjective experiment emulates a simplified version of the voice-casting process using source language VTK from our content along with foreign language dubbing tests across multiple different languages for the same characters. We aim to establish a subjective testing framework for future studies in the area as well as present preliminary results to the fidelity of the speaker embedding approach for voice casting. We begin by describing an automated voice-casting workflow that we developed and used for evaluation. Then, we provide a detailed overview of the subjective voice-casting study we conducted, followed by the results and discussion of the results.

Automated Voice-Casting Tool

The automated voice-casting tool used for evaluation consists of two main steps, as shown in **Fig. 1**. The first step is to compute high-dimensional embeddings for each voiceover artist using a speaker encoder model trained on a speaker-verification task. The second step is to compute similarity scores between embeddings of source language reference speakers and target language voiceover artists to generate rankings.

Speaker Encoder Model

To enable an automated workflow as described earlier, our deep learning encoder model needs to satisfy two main requirements. First, it should be able to generate embeddings for unseen speakers during training (zero-shot). Second, it should also be able to generate embeddings from a small duration of audio (typically under 60 sec for dubbing VTKs). The encoder model we used to generate embeddings for multilanguage voiceover artists is shown to satisfy both requirements listed above. It is a three-layer long short-term memory (LSTM) with 768 hidden nodes followed by a projection layer of 256 units as described in Ref. 9. It was trained on a speaker-verification task using generalized end-to-end loss.[6] The inputs to the model are 40-channel log-mel spectrograms—acoustic time–frequency representations of audio—with a 25-ms window width and a 10-ms step. Mel spectrograms are commonly used as inputs for end-to-end encoder models.[10,11] The log-mel spectrograms were processed using librosa Python library. The model was trained using open data sets LibriSpeech-Other, VoxCeleb1, and VoxCeleb2.

Ranking System

The output of our encoder is a 256-dimensional embedding for each speaker, which can then be compared to other speakers using a similarity score generated using cosine similarity. We can compute similarity scores between embeddings of source language reference talents and target language voiceover artists. The highest similarity scores would correspond to the highest perceptual similarity for a source–target language pair. We then directly utilize similarity scores to generate rankings for target language dubbing voiceover artists in comparison to a source language reference talent. The highest similarity score would be considered as the top-ranking voiceover artist for a source–target language pair. These rankings and scores can then be shared with the experts to assist in their decision-making process.

Advantages

There are opportunities for an automated system as proposed earlier to augment the existing voice-casting workflow. As mentioned before, the process of voice casting today is mostly a subjective process. Having an objective quantitative metric of perceptual similarity such as the similarity scores described could prove to be a useful additional data point to aid the casting decisions. Furthermore, evaluations of auditions are

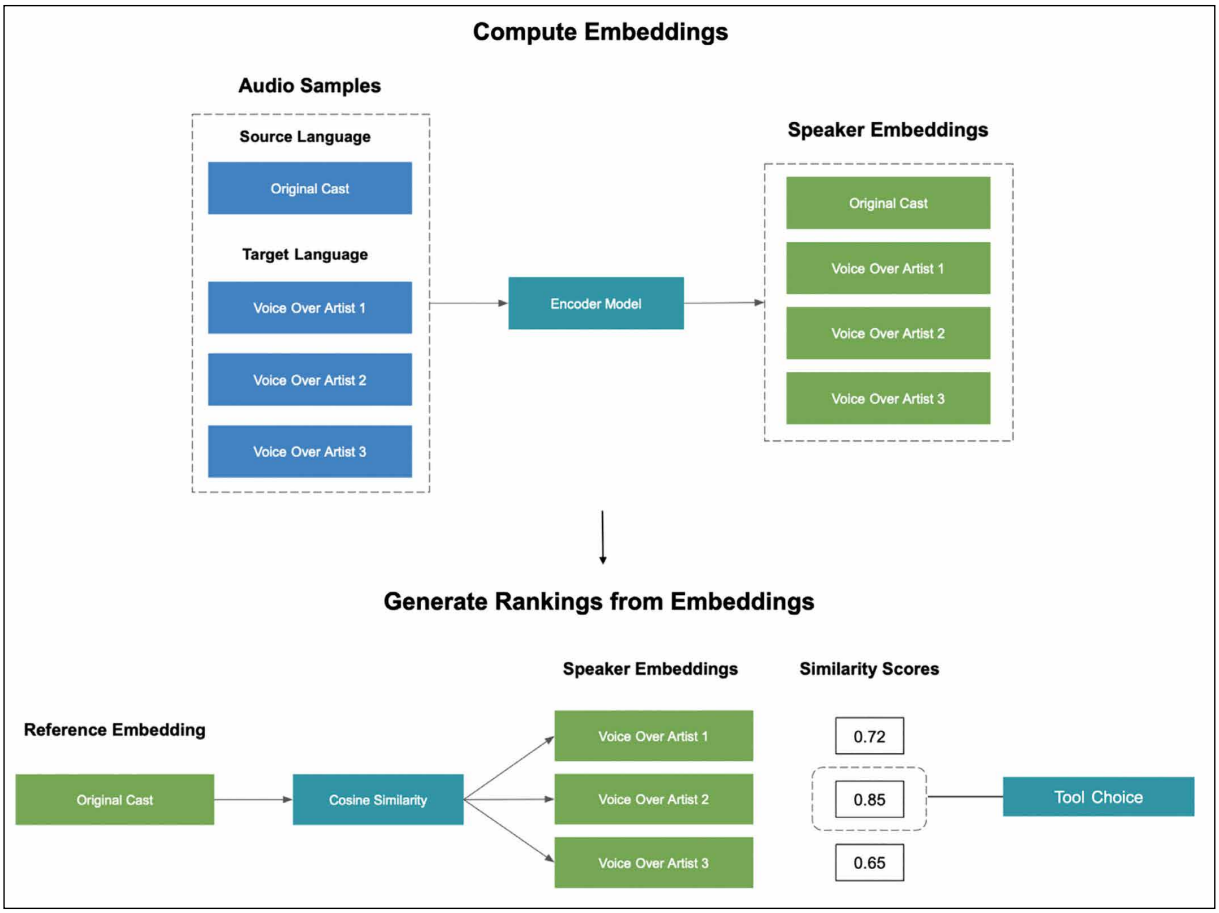


FIGURE 1. Automated voice-casting tool overview.

time-consuming in nature. They often involve multiple playthroughs and careful consideration for each audition, which when casting for high volumes of content is a considerable time commitment. An automated tool as proposed can be employed to make tertiary evaluations. The tool can evaluate a large number of auditions and filter down to the top few candidates for each character for the experts to evaluate. This could enable auditions with a much larger and diverse pool of talents. The tool thus has the potential to make the workflow more scalable and objective in nature.

Subjective Voice-Casting Study

Our goal for the subjective study was to emulate a simplified version of the real voice-casting process carried out during content localization. We could then evaluate as well as compare results with the automated voice-casting tool. In practice, voice-casting experts are provided VTK for source (in our case, English) as well as target languages (e.g., French, Spanish, and Portuguese). The source language VTK consists of audio samples for each character in a specific content. These audio samples are typically around 60 sec in duration consisting of a few different dialogs acted out by the original cast playing a certain character in the source language content. The target language VTK consists of audio samples from multiple voiceover artists auditioning for dubbing a character in the content. The voiceover artists typically perform the same dialogs from the source language audio sample in the target language. The task of voice-casting experts is to evaluate the perceptual similarity of the voices of these target language voiceover artists to the original source language cast and select the most similar voices for each character.

We conducted a study using actual dubbing VTKs from a Warner Bros. television series consisting of audio samples for four characters (one female and three males). The source language was American English and the target languages were Latin American Spanish, Brazilian Portuguese, Italian, and Hungarian. **Table 1** shows the number of voiceover artists auditioning for dubbing each character. In addition to the VTK, we also had access to the voice-casting decisions made by the casting experts. The casting decisions would not only help us directly evaluate the accuracy of the tool, but also peek into the overall subjectivity of perceptual similarity between voices across different languages. Typically, the voice-casting process for a content is carried out by one expert. Experimenting with a bigger sample size would help us understand whether there exists a constancy in the perceptual similarity observed by humans in general.

TABLE 1. Number of voiceover artists for each language–character pair.

Language	Number of characters	Number of voiceover artists per character
Latin American Spanish	4	3
Brazilian Portuguese	4	2
Hungarian	3	3
Italian	3	3 (4 for one character)

Subjective Test Procedure

The tests were conducted online using Microsoft forms. There were a total of 56 participants. All participants were instructed to wear headphones for the duration of the test. In addition, we conducted a frequency response test at predefined volume conditions to ensure auditory fidelity of the setup as well as participants' hearing. Participants were asked to record lowest and highest frequencies they were able to hear through their headphones. The structure of the test shown in **Fig. 2** was as follows.

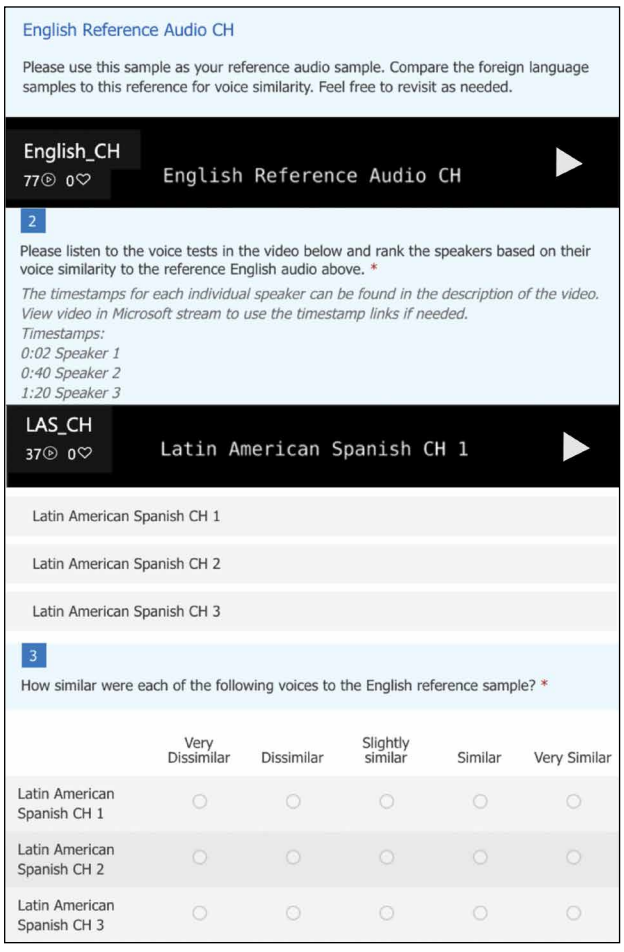


FIGURE 2. Snippet from the Microsoft Forms Survey used for the subjective test. Participants were first shown the English reference audio, then the dubbing tests in target language. The first question was to rank the speakers based on their similarity to reference sample (Question 2) followed by a similarity score (Question 3).

We first had an American English reference audio sample from the VTK, followed by an audio sample containing dubbing tests from multiple voiceover artists for the same character. Tests following the same format were sent out for each target language and each character. Participants could reference, pause, and play both the source as well as the target language audio samples at their will to help with their evaluations. They then had to answer three main objective evaluation questions shown in **Fig. 2**, which included ranking the voiceover artists based on their similarity to the English reference audio, scoring each voiceover artist based on similarity to reference as well as scoring their confidence in their ranking/similarity scoring.

Ranking Candidates

The first question was ranking the voiceover artists based on their similarity to the English reference audio sample. Participants would choose a ranking from 1 to n, where n is the number of voiceover artists auditioning. Rank 1 would be their top casting choice and n their least favorite. This emulates a simplified version of how voice casting is done today.

Similarity Scores

In addition to ranking the candidates, participants were asked to assign a similarity score to each voiceover artist. The similarity scoring followed a five-point scale as follows: (–2) Very Dissimilar, (–1) dissimilar, (0) fairly similar, (1) similar, and (2) very similar. Participants were also asked to rank their confidence in their scoring using a similar five-point scale as follows: (–2) not confident, very challenging, (–1) challenging, (0) confident, but still challenging, (1) confident, and (2) very confident.

Automated Tool Test Procedure

The same VTKs were used as input to the automated tool for evaluation. The automated tool used the source language speaker embeddings as reference embeddings and then computed the similarity scores for the target language voiceover artist per character. The highest scoring voiceover artists were then recorded as the chosen voiceover artist for comparison with the subjective study decisions as well as the casting-expert decisions.

Results

Statistical Significance for Subjective Study Rankings

Before evaluating and comparing the rankings from the subjective experiment, it is important to establish statistical significance for the rankings from the study. A ranking would only be considered statistically significant (SS) if it satisfies the condition that the margin of error is less than the difference in proportions for the top two contenders for the top-ranking voiceover artists. This equation is outlined in Ref. 12 and shown in Eq. 1

$$CI(p1-p2) = t-score \star \sqrt{\frac{(p1+p2)-(p1-p2)^2}{n-1}}. \quad (1)$$

We switch out the z-score in the equation from Ref. 12 with t-score for 95% confidence intervals due to our low sample sizes for each individual study. In the equation, p1 and p2 are proportions for the top-ranking voiceover artist. Interestingly, we found that just under 43% of the results had SS rankings. The low statistical significance rates could be attributed to either the commendable competency of the dubbing talent pool used for the test or the subjectivity/variability in perceptual similarity between voices. The latter would point to the inherent difficulty in the voice-casting process and would deem having an objective guiding metric useful. These results are discussed in detail in the "Discussion" section.

Automated Tool Results Evaluation

To evaluate the accuracy of the tool, we used the match rate as the main metric. Matches were considered as every language–character pair for which the automated tool predicted the same highest ranking (highest similarity score) voiceover artist as the study participants did. This would be the appropriate metric for the use case of voice casting in Eq. 2.

$$\text{Match rate}(\%) = \left(\frac{\text{Number of matches}}{\text{Number of matches} + \text{Number of nonmatches}}\right) \times 100. \quad (2)$$

Match rates for all evaluations are shown in **Fig. 3**. To begin with, we evaluated the performance of the automated casting tool for only the SS rankings from the study. We achieved a remarkable match rate of 100% for this case. This meant that every time there was a significant consensus between our participants that a voiceover artist was more acoustically similar to the original cast member than the other artists, the automated tool ranked the same artist as most similar.

Due to the low number of SS rankings, we decided to consider other rankings that showed a clear majority ranking using the criterion outlined in Eq. 3

$$\begin{aligned} &\text{if } n: \text{ even, majority} > \frac{n}{2}+1 \\ &\text{if } n: \text{odd, majority} > \frac{n+1}{2}. \end{aligned} \quad (3)$$

The variable n in the criterion signifies the sample size. Using the above criterion, 71.4% of the rankings showed a clear majority. For this case of rankings with majority votes (MVs), we observed a match rate of 77.8%.

Next, we decided to compare both the automated tool rankings as well as the study rankings to the casting-expert decisions. For the case of SS rankings, we observed only one mismatch between the expert decisions and the automated tool as well as the study rankings. Including all rankings that had MV, we observed a 77.8% match rate between the automated tool and the expert rankings

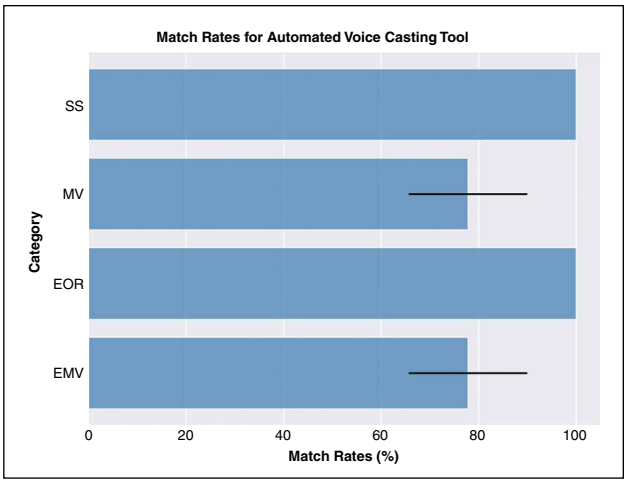


FIGURE 3. Match rates for automated voice-casting tool with 95% confidence intervals. The categories of match rates are as follows. SS: statistically significant rankings; MV: rankings with majority votes; EOR: either study rankings or expert decision matches; EMV: expert decisions for rankings with MVs.

(EMV). On further analysis of the mismatches, we found an interesting observation. For the MV case, we computed an either–or match rate. The either–or match rate (EOR) considered a match if the automated tool matched with either the study rankings or the expert rankings or both. We observed a 100% either–or match rate for MV cases, which meant that everytime the automated tool ranking mismatched with the study rankings, it matched with the expert ranking and vice versa. For MV cases, we observed a match rate of only 55% between the study rankings and the expert decisions. This is discussed further in the "Discussion" section.

Discussion

First, we explore the interesting observations of the low rate of statistical significance for our study rankings as well as the low match rate between the top-ranking voiceover artists for the study versus the expert-casting decisions. As mentioned before, we found just under 43% of the rankings from the study to be SS. For the match rate between the study rankings and the expert-casting decisions, we saw almost a 100% match rate for the SS rankings but only a 55% match rate for all rankings that satisified a MV. Our initial instinct was to give more weight to the expert-casting decisions owing to their experience, expertise, and character understanding. However, the end-users for our content are the viewers and thus it is just as important to factor in the similarity perceived by the study audience as it is to factor in casting expertise. Investigating the mean perceptual similarity scores with 95% confidence interval for the insignificant cases shown in **Fig. 4** added some clarity to the observation.

All the language–character pairs except one (Hungarian_LI, which had proportions that were borderline significant) showed an overlap of confidence intervals.

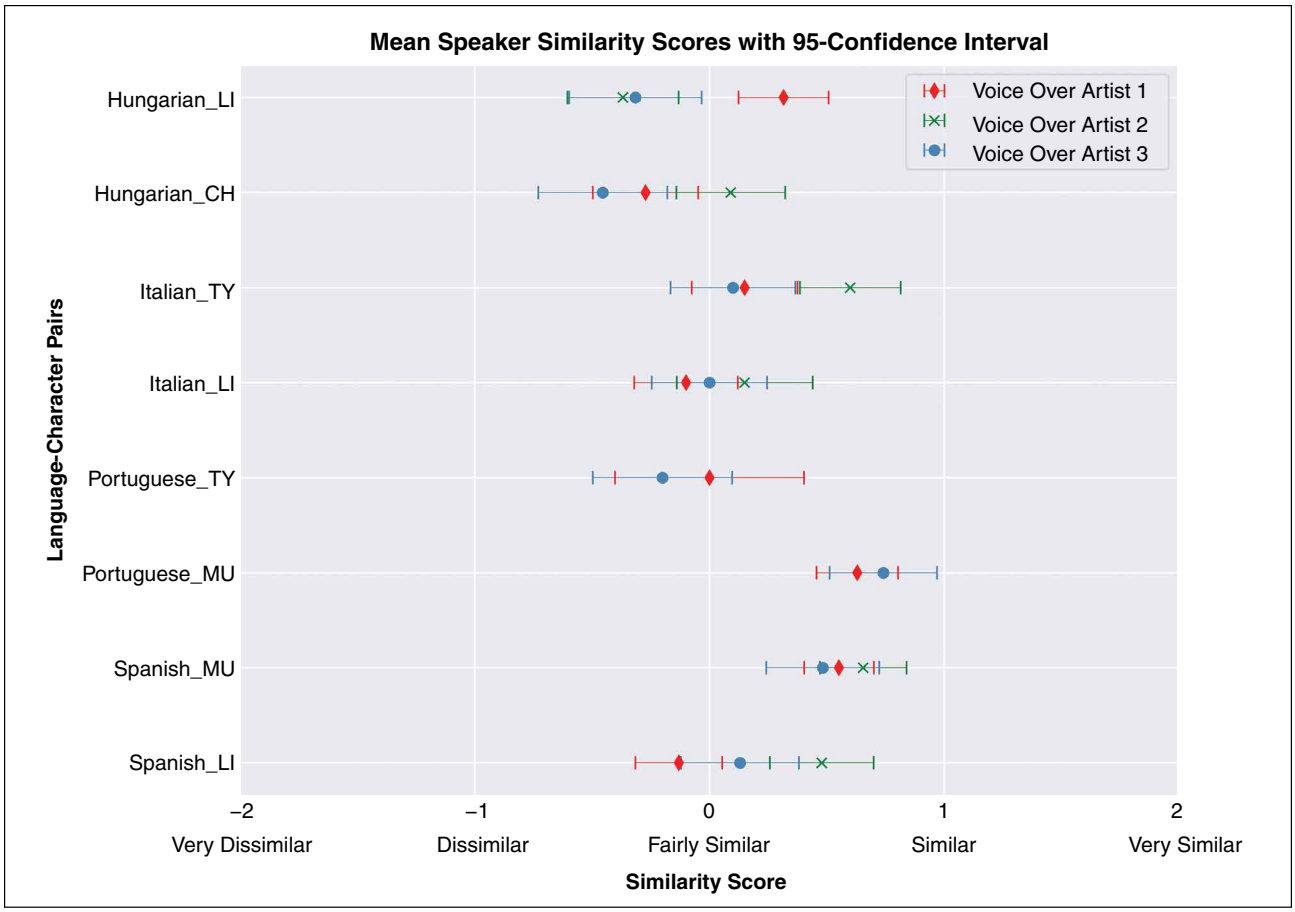


FIGURE 4. Mean speaker similarity ratings from subjective study with 95% confidence intervals for statistically nonsignificant rankings. The characters from the television series used for the test are abbreviated as CH, LI, MU, and TY on the language–character pair axis.

With this, we can conclude that the multiple voiceover artists for those language–character pairs perceptually were roughly in the same range of perceptual similarity to the reference speaker (original cast). In such cases, rankings would be subjective since it would be inherently difficult to differentiate between multiple voiceover artists based on perceptual similarity to the reference. In such cases, where the talent pool is perceptually around the same level of similarity to the original cast, it is plausible that experts defer to meta-information (such as talent history, character notes, etc.) that was not available to us or the study participants to make their decisions. Either way, such cases exemplify the need for a trustworthy objective metric to guide the voice-casting process and alleviate some of the subjectivity in the process today.

Next, we delve into the match rates for the rankings from the automated tool versus the study audience as well as the casting experts. It should be noted that the data corpus spanning just one content as well as four languages is not enough to draw concrete conclusions to the dexterity of the proposed automated workflow. However, the match rate of 100% for the case of SS, 77.8% for MV, EMV, and 100% EOR (either expert decisions or study audience rankings) for the MV case are promising. At the very least, these results show the promise of the tried approach of using similarity scores from neural network embeddings trained for the task of text-independent speaker verification (TI-SV) in the domain of voice casting and prove that they warrant further comprehensive investigation with a larger data corpus.

Conclusion

In this article, we explored the viability of using similarity scores computed using neural network speaker embeddings trained for TI-SV for automating the voice-casting process. We proposed an automated workflow for the same and attempted to evaluate its dexterity using both a subjective study and casting expert decisions. We conducted a subjective study accurately emulating a simplified voice-casting process using actual VTKs from a Warner Bros. television series. Our aim was to establish a framework for studies in this domain that can be repeated for future work. Analyzing the results from the study uncovered some of the inherent subjectivity in the task of voice casting,

which should be investigated further in future studies. The subjectivity accentuates the benefit of having more objective metrics to guide the voice-casting process. Due to the limited size of our test corpus, we were only able to show preliminary results to the dexterity of the automated workflow. Nevertheless, the results were promising and show that such a technique warrants further comprehensive investigation. In future studies, we would like to work with a larger data corpus, conduct studies with a larger sample size, as well as collect more objective metrics from casting experts to shed light on their current evaluation techniques.

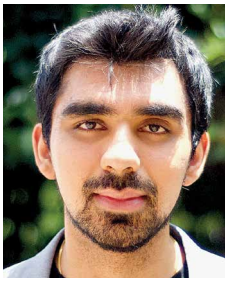
Acknowledgment

The authors would like to thank the Content Localization Team at WarnerMedia for providing the VTK as well as data used in this study, with special thanks to Grant Duncan, Amy White, and Jasbir Tang. The authors also thank Darren Kim for his help with data analysis as well as all the WarnerMedia employees who participated in the subjective study.

References

1. N. Obin, A. Roebel, and G. Bachman, "On Automatic Voice Casting for Expressive Speech: Speaker Recognition vs. Speech Classification," *Proc. IEEE Int. Conf. Acoustics Speech Signal Process, (ICASSP)*, Florence, Italy, pp. 950–954, 2014, doi: 10.1109/ICASSP.2014.6853737.
2. N. Obin and A. Roebel, "Similarity Search of Acted Voices for Automatic Voice Casting," *IEEE/ACM Trans. Audio Speech Language Process.* 24(9):1642–1651, Sep. 2016, doi: 10.1109/TASLP.2016.2580302.
3. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digit. Signal Process.*, 10:19–41, Jan. 2000.
4. W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support Vector Machines Using GMM Supervectors for Speaker Verification," *IEEE Signal Process. Lett.* 13(5):308–311, May 2006, doi: 10.1109/LSP.2006.870086.
5. A. Gresse, M. Rouvier, R. Dufour, V. Labatut, and J.-F. Bonastre, "Acoustic Pairing of Original and Dubbed Voices in the Context of Video Game Localization," *Proc. Interspeech*, Stockholm, Sweden, pp. 2839–2843, 2017, doi: 10.21437/Interspeech.2017-1311. hal-01572151.
6. L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized End-to-End Loss for Speaker Verification," *Proc. IEEE Int. Conf. Acoustics Speech Signal Process. (ICASSP)*, pp. 4879–4883, 2018, doi: 10.1109/icassp.2018.8462665.
7. E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification," *Proc. IEEE Int. Conf. Acoustics Speech Signal Process. (ICASSP)*, Florence, Italy, pp. 4052–4056, 2014, doi: 10.1109/ICASSP.2014.6854363.
8. D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep Neural Network-Based Speaker Embeddings for End-to-End Speaker Verification," *Proc. IEEE Spoken Language Technol. Workshop* (SLT), San Diego, CA, pp. 165–170, 2016, doi: 10.1109/SLT.2016.7846260.
9. Y. Jia et al. (Jan. 2, 2019). Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis. Retrieved Jul. 9, 2020, from https://arxiv.org/abs/1806.04558
10. L. Deng et al. "Binary Coding of Speech Spectrograms Using a Deep Auto-Encoder," *Proc. INTERSPEECH-2010*, pp. 1692–1695, 2010.
11. Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop," *Proc. Int. Conf. Learning Representations (ICLR)*, 2018.
12. A. J. Scott and G. A. F. Seber, "Difference of Proportions From the Same Survey," *Am. Stat.*, 37(4a):319–320, 1983, doi: 10.1080/00031305.1983.10483130.

About the Authors



Aansh Malik is an AI engineer in the Emerging and Creative Technologies Team at WarnerMedia. His work focuses on leveraging advancements in AI to enable more interactive storytelling, facilitate experiences in immersive mediums such as AR/VR, and augmenting current workflows in the media production lifecycle. His projects are a confluence of fields including reinforcement learning in game engines, computer vision, recurrent neural networks for audio processing, among others. He has a BS degree in electrical engineering with a specialization in machine learning from the University of California at San Diego (UCSD), La Jolla, CA.

Ha Nguyen is the director of emerging technology at WarnerMedia's Emerging and Creative Technologies, Burbank, California. With a background in computer science, she has been with Warner Bros., now WarnerMedia, for 16 years. Starting out as a software engineer for mobile, back-end operations, and web, she ventured into digital and physical media, its application, post-production, and distribution services. She then switched gears to work in the emerging technology arena, where she helped to launch all WarnerMedia titles in ultra-high-definition media format. Nguyen brings mixed reality, artificial intelligence, and future storytelling concepts to the studio, and participates in research and development work for next-generation media content creation.

Presented at the SMPTE 2020 Virtual Conference, 10–12 November 2020.