The AIIDE-17 Workshop on
Experimental AI in Games
WS-17-19

Deep Learning for Speech Accent Detection in Videogames

Astrid Ensslin, Tejasvi Goorimoorthee
Humanities Computing
University of Alberta
Edmonton, AB
{ensslin | tejasvi}@ualberta.ca

Shelby Carleton
English and Film Studies
University of Alberta
Edmonton, AB
scarleto@ualberta.ca

Vadim Bulitko, Sergio Poo Hernandez
Computing Science
University of Alberta
Edmonton, AB
{bulitko | pooherna}@ualberta.ca

Abstract

In video games, a wide range of characters make up the world players inhabit. These characters, NPCs, have traits, such as their appearance and speech accent, that determine certain things about them, including moral inclination, levels of trustworthiness, social class, levels of education, and ethnic background. But what does an accent say about a character in a video game? We use deep learning to train a neural network to detect speech accents and establish the degree to which machines can be used to recognize these accents. This research aims to help sociolinguists and discourse analysts establish critical study and content analytical findings for instance about stereotypical uses of speech accents, to better analyze who has what accent in video games, and what kind of language ideologies and social value judgments the use of accents in games construct and perpetuate. This paper presents the results of the first deep learning experiments, which were conducted on Standard North American, British Received Pronunciation, and Spanish English. We discuss our methodological considerations and some early deep learning results, which show relatively low levels of accuracy (61%). We discuss possibilities of improving our method, and of enriching our training datasets.

1 Introduction

This paper is part of a larger research project called "Speech Accents in Games" that looks at how selective uses and distributions of speech accents in a range of narrative video games may cast light on underlying language ideologies, and specifically accent attitudinal evaluations (e.g., in terms of prestige or social attractiveness). Attitudes towards speech accents (such as "Oh I love this Italian accent," or "Why is the villain always British?," (Poos 2013)) can contribute to linguicist behavior and politics. Much like racism and sexism, linguicism refers to popularly held, stereotypical views about other peoples use of language in speech and writing. These views tend to be judgmental and can cause considerable harm in those they are directed against. For example, negative biases towards people's regional accents may lead to exclusion or side-lining on the job market and in mainstream media representations, or even to downright bullying.

Certain characters in video games are associated with specific attributes that designates a character to be, for example, a hero or villain, or impoverished or rich. In *Star Wars: Knights of the Old Republic*, for example, the color of the lightsaber carried by the player-character designates an allegiance to either the Sith or Jedi (BioWare 2003).

Extending this idea that certain characteristics can denote certain ideas beyond just outward appearance or material artifacts, it has been established that character voices, intentionally or unintentionally, are often used to say something about a video game character, their moral inclinations, levels of education, as well as ethnic and class background. Learning what accents are used where in the medium of video games will give a more nuanced picture of how friend and foe, as well as more hybrid, dynamic, and rounded character roles may be framed phonetically - through allocation of linguistic accents - and how these vocal attributes may tie in (or not) with other semiotic modes such as visuals (facial features; body language) and written language (e.g., dialog).

When giving accents to certain characters, it becomes important to consider the possible implications of a certain accent connecting to a particular character. For example, if a British accent is often assigned to a villain, British accents will become associated with evil characters, thus iconizing British varieties as symptoms of undesirable moral inclinations. By learning about certain accent patterns and studying them systematically, we hope to encourage independent, experimental game designers in particular to aim for more diverse and egalitarian representations of linguistic realism.

2 Problem Formulation

The use of accents in video games can be just as ideologically charged as any other element of audiovisual representation, such as color of skin, type of clothing, and accompanying soundtrack. These semiotic choices can have the intent of being neutral, but sometimes, they can also communi-

cate certain assumptions and ideologies about what players are supposed to make and have about characters in a game. The distinction between a player-character and a non-player character is medium-specific, and in most narrative games players tend to encounter a wealth of mostly flat, scripted non-player characters that function as quest givers, navigation aids, clues, or have seemingly no specific purpose at all. For the most part, they fill a purely functional role with few idiosyncratic personalities, and their language tends to default to simple sentences, grammar, and phonology. In relation to speech accents, this means their use of speech generates a so-called matrix of predominant accent use, such as an RP matrix (if most characters speak with Received Pronunciation), or a Standard North American matrix. Accents that fall outside these matrices are, as a result, marked in that they represent deviations from the norm, and this gives rise to semiotic processes of othering: exposing specific individuals as different from the norm and therefore potentially threatening or at least suspicious to everyone else.

Generally speaking, video game design uses conventional and unconventional semantic oppositions to construct simplistic Manichean binaries of morally good or bad behavior and to cluster them together with aesthetic features such as bright/positive versus dark/negative. Players are subjected to the use of semiotic signals coded into the user interface, and therefore led to adopt the binary persuasive logic of the game mechanics. In relation to speech accents, this may result in artificially constructed binary concepts, for example, of Standard North American used in positively connoted characters, versus Arabic or other foreign accents used in negatively connoted characters. This again can lead to a perpetuation of stereotypical, hegemonic thinking about "the dominant culture and prevailing power relationships" (Schniedewind and Davidson 2000).

Narrative games feature a host of non-player characters, and to study their accents comprehensively and systematically, we are proposing a method that will help us accelerate processes of accent recognition, as well as improve accuracy of recognition. To do so, we first need to train the computer to recognize speech accents while being robust to background music, using an existing speech accent database, the Speech Accent Archive (Weinberger 2015).

3 Related Work

Early sociolinguistic work into speech accents in games looked at how conventional and unconventional semiotic oppositions are conflated in multimodal video game character representations, combining aspects of character speech with other types of semiotic information, such as visual appearance, body language, and musical soundtrack (Ensslin 2010). Examples include the pairing of moral binaries (good and bad) with artificial opposites like Received Pronunciation versus New York English, and by looking at how Pax Americana (American hegemonic superiority; (Bayard et al. 2001)) is embedded in and iconized by the voices of heroic characters (Irvine and Gal 2009; Lippi-Green 2012). This research is complemented by Brice (2011), who studies Bioware's *Dragon Age* with respect to how a multitude of American accents surrounding the player helps render the American accents invisible so the player may focus on other things. However, the opposite becomes true within Antivan conflicts, where Spanish accents are prominent, highlighting a difference of location and social belonging between two groups which seems to be the primary focus. A further study by Ensslin (2011) examined how speech accents can be used (a) as tools of othering undesirable characters, and (b) as mnemonic devices, generating intertextual links to other elements of popular culture, thus increasing levels of immersion and entertainment. This early research was largely centered around a narrow set of narrative 3D games (*Black and White 2* (Lionhead Studios 2005); *Return to Castle Wolfenstein* (id Software and Grey Matter Interactive 2001); Aion (NCsoft 2008); *Fable* (Lionhead Studios 2004); *Wizard 101* (KingsIsle Entertainment 2008); and *Dragon Age: Inquisition* (BioWare 2014)).

In the field of machine learning, convolutional neural networks (CNN) have previously been used for spectrographic analysis of sound data for music classification (Costa, Oliveira, and Silla 2017) and for recognizing aspects of human speech. With respect to the latter, Huang et al. (2014) report on a study training CNN for affect-salient features for Speech Emotion Recognition, with positive results in recognition performance in complex scenes, including speaker and environment distortion. Using a deep CNN model, Badshah et al. (2017) report better performance on Speech Emotion Recognition than when fine-tuning a pre-trained AlexNet model. The research of Espi et al. (2015) into acoustic event detection emphasizes the importance and feasibility of local feature extraction in detecting and classifying non-speech acoustic events occurring, for example, in conversation scenes. As opposed to CNN, recurrent neural networks (RNN) as a possible machine learning method for speech recognition has so far been disappointing (Graves, Mohamed, and Hinton 2013).

In the area of accented speech recognition, Hautamäki et al. (2015) explain that "foreign accent variation is a nuisance factor that negatively affects automatic speech, speaker and language recognition systems." They investigate the use of two deep neural networks with nodes for modelling the speech attributes manner and place of articulation, demonstrating the effectiveness of CNNs for attribute classification and foreign accent recognition.

4 Our Approach

We began by taking an existing convolutional network and training it as follows. Initial training was executed using recordings from the Speech Accent Archive. Following training, we attempted to classify accents from recordings obtained from captured videos of *Dragon Age: Origins* (BioWare 2009) (Figure 1). These captured files were separately labeled by two proficient (one native and one near-native) English speakers. Each speaker listened to each audio file, and gave it a label for specific accents of native and foreign English—including Spanish, British, American, and French. In this process, the two labelers compared their results and debated any dissimilar labels to finalize the *Dragon Age* accent database, where each character was

given a distinct accent label to be compared against the network's own classification.

The *Dragon Age* series was chosen as it re-appropriates accent dynamics for the assumed North American player. *Dragon Age: Inquisition* (the third installment of the series) for example embeds various European accents but attributes them to specific character races in stereotypical ways: the Dalish speak with Irish accents, Antivans [anteevans] with Spanish accents, Orlesians with French accents, and Nevarrans with Russian accents. City elves, dwarves, and the Qunari, contrastively, have American accents, which still comes across as unmarked despite the underlying RP matrix. The Fereldans, finally, have different British accents. This wide range of races and accents in *Dragon Age* are hoped to yield a variety of results when training the accent data with a network.

Figure 1: *Dragon Age: Origins.* Player talking with Duncan who has a British accent.

Using AlexNet (Krizhevsky, Sutskever, and Hinton 2012) as our network, we adapted it to our task by removing the last layer and replacing it with a fully connected layer of size three: the number of classes we are trying to classify. However, as AlexNet accepts images as its input, we created a visual representation of the audio files by converting them into spectrograms.

Each spectrogram consists of four different image quadrants (Figure 2). The first quadrant is the spectrogram with frequency and amplitude as linear parameters. The second quadrant is linear frequency and we apply a logarithmic function to the amplitude. In the third quadrant a logarithmic function is applied to the frequency and linear amplitude. In the fourth quadrant, a logarithmic function is applied to both amplitude and frequency. Each of the changes applied to the parameters is done with the hopes of providing more information about the audio to the network. By applying the logarithmic function to the amplitude, we are increasing the resolution for the frequency filters with low amplitude, but we are decreasing the resolution for the frequency filters with high amplitude. Similarly, when applying the function to the frequency filters, we are increasing the resolution provided by the lower frequencies, at the expense of the higher ones. The colors are determined by the use of a color map: the lower the amplitude the colder the color (blue) and the higher amplitude the hotter the color (red).

Audio file length must also be taken into account. Each audio file in the archive is approximately 20 seconds long, and creates a large, dense spectrogram. These image proportions may lead to a loss of information if the spectrogram needs to be resized to a smaller image. Therefore, in order to prevent the loss of information, each audio file is divided into smaller sized files and spectrograms are created for each of these segments.

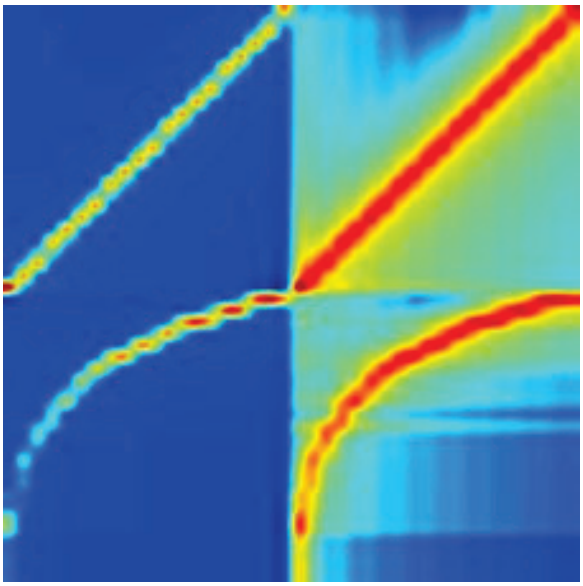

Figure 2: A 2×2 spectrogram of audio with increasing pitch. The first quadrant (top left) shows spectrogram of audio with linear parameters. The second quadrant (top right) shows a spectrogram with linear frequency and logarithmic amplitude. The third quadrant (bottom left) shows a spectrogram with logarithmic frequency and linear amplitude. The fourth quadrant (bottom right) shows a spectrogram with logarithmic parameters.

5 Implementation and Empirical Evaluation

We aim to train a network to be able to differentiate between American (Figure 3), British, and Spanish (Figure 4) accents. These languages were selected because they are the most prominent in the *Dragon Age* series.

5.1 Training on Speech Archive Data

We divide the audio files in 3 second segments and create a spectrogram for each segment. The spectrograms are divided into a training set and testing set, used to train and validate the network. We use the Matlab neural network toolbox to train the network. In order to determine the best value for each parameter to use with the network, experiments are run with a range of values before selecting the set that gives the best test results. We are interested in four parameters, two related to the network: number of epochs and the size of the batch used; and two related to the size of the spectrograms: number of time windows and number of frequency filters.

When dividing the data into training and test sets, all spec-

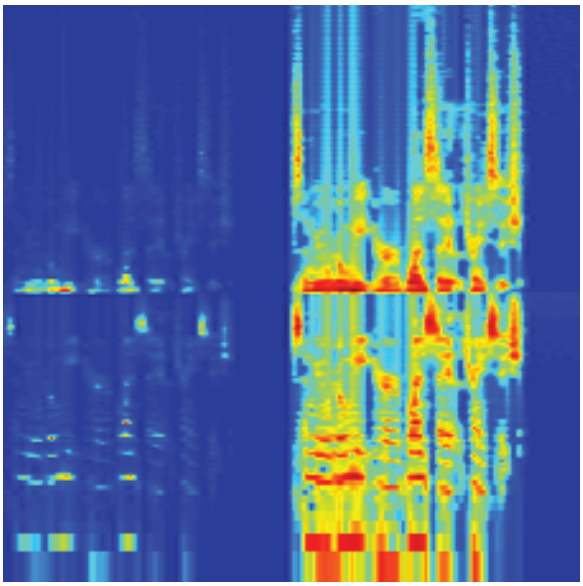

Figure 3: A sample spectrogram of a speaker with American accent, used in training the neural network.

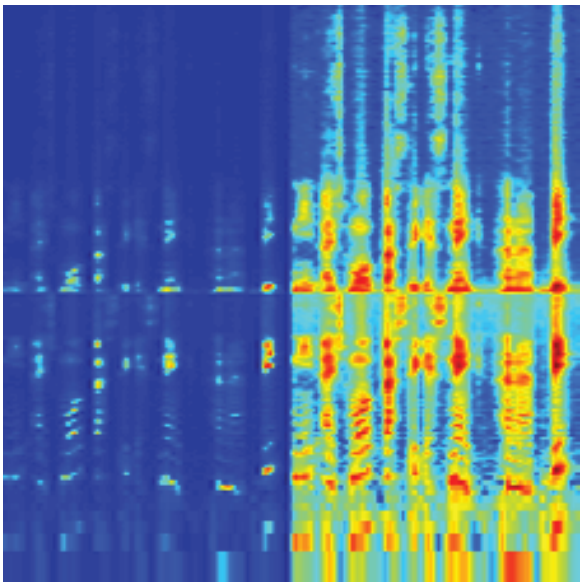

Figure 4: A sample spectrogram of a speaker with Spanish accent, used in training the neural network.

trograms corresponding to the same audio file are ensured to be in one of the sets. The input for AlexNet is a 227×227 image, so we resized the spectrograms to these dimensions regardless of their original size. The ranges tested for each parameter are: 10, 50, 100 and 200 epochs, 3, 5, 10, 50 and 100 batch sizes, 10, 25, 50, 75 and 113 number of windows and frequency filters. We divided the audio files into 1, 3 and 5 seconds segments, and of these three the tests, the 3 second segments produced the highest accuracy. We train different networks with a combination of these four parameters to determine the values best suited for our task.

To train the networks we used the Speech Accent Archive audio files. For the experiments used to determine the best parameters for our task, four trials are run per parameter combination. On each trial we randomly divided the 67 audio files per class (US, UK and SP) into 75% used for training and 25% used for testing. The training accuracy was averaged over the four trials. For audio segments of 3 seconds, we determined the following parameters as the best: 50 epochs, 5 images per batch, 113 frequency filters and 50 time windows. With these parameters the network had an average test accuracy of 61%.

Once we had the parameters we ran a test to create a confusion matrix to determine which classes the network is confusing. For this experiment we ran eight trials. Table 1 shows the results of this experiment. Each column in the matrix represents how the network classifies the files of that label. For example for British the network classified 61% of the files as British, 20% as American and 19% as Spanish, showing us which labels are confused the most.

5.2 Classification of Video Game Files

Once AlexNet was trained, we froze its weights and applied it to audio files captured from *Dragon Age: Origins*. The 17 audio files were labeled as American and British yet the network was trained on American, British and Spanish subset of the Speech Accent Archive. There were a total of 17 *Dragon Age: Origins* files which were subjected to the segment division and spectrogram construction process described in the previous section. AlexNet labeled most of them as Spanish, achieving the accuracy of only 9%. The unexpectedly poor performances prompted to re-train AlexNet on American and British accents only. Using the same training parameters and 134 audio files (67 with American accent and 67 with British accent) we now achieved a higher test accuracy of 75% with the confusion matrix found in Table 2.

Table 1: The confusion matrix: British, American, Spanish.

	Actual		
Recognized	British	American	Spanish
British	61%	20%	12%
American	20%	68%	10%
Spanish	19%	12%	78%

Table 2: The confusion matrix: British and American.

	Actual	
Recognized	British	American
British	73%	23%
American	27%	77%

We used the same 17 audio files from *Dragon Age: Origins* to determine how reliable the network became with the new change. The network achieved the accuracy of 52.7%, which exceeds the original 9% accuracy.

Since the audio files used to train the network had no background noise we used 10 new audio files with no background music captured from *Dragon Age: Origins* to determine if this had an effect on the network's accuracy. The network correctly recognized 6 out of 10 files (60%).

6 Current Challenges and Future Work

The empirical results suggest a few questions for follow-up studies. First, how authentic are accents in video games, done by voice actors who act them out and may or may not be the native speakers of the target accent? Second, how limiting was the training on the Speech Archive data where each speaker reads the same sentence, clearly and unemotionally? This is in contrast to voice performance in video games where the text, tempo and emotional color can vary dramatically based on the context. Third, what is the best way to determine correct labels for accents in video games? How do we establish that an actor voice-acting an elf used a British accent? Some "fantasy" accents may not even fall clearly in real-life accent classes.

Future goals of this project are to examine, as systematically and comprehensively as resources will allow, how different accents of English are distributed, combined with other semiotic sources for multimodal meaning, how they are functionalized politically in the game world, and what kind of language and specifically accent attitudinal ideologies they might perpetuate in this way. The main deliverables of the project are a large and conceptually growing database of speech accent samples (video and audio files), including metadata, and a number of methodological, technical, and critical papers. The main analytical approaches will be (a) qualitative text analysis of individual game franchises, used as case studies, such as *Dragon Age*, (b) content analysis, to get a sense of distribution and spread of standard versus non-standard varieties across major international video game franchises, and (c) further machine learning tests, comparing audio files from video games data with existing speech accent databases and continuing to establish the extent to which computers can identify accents from videoclips.

Finally, it would be of interest to attempt to train a neural network to automatically label audio files with non-accent attributes. For instance, a network that can detect authenticity of anger expressed by a voice actor may be useful to a video-game developer in preliminary filtering of audition submission for a character.

7 Conclusions

In this paper we presented a preliminary study on training a deep neural network to automatically label audio files with accents. We trained AlexNet on the Speech Accent Archive data and then applied it to audio files captured from a video game. The first results appear promising and open several exciting avenues for follow-up work.

8 Acknowledgments

We appreciate the support from Kule Institute for Advanced Study (KIAS), the Social Sciences and Humanities Council of Canada (SSHRC) via the Refiguring Innovation in Games (ReFiG) project, the Alberta Conservation Association, the Alberta Biodiversity Monitoring Institute, and Nvidia.

References

Badshah, A. M.; Ahmad, J.; Rahim, N.; and Baik, S. W. 2017. Speech emotion recognition from spectrograms with deep convolutional neural network. In *Proceedings of 2017 International Conference on Platform Technology and Service (PlatCon)*, 1–5.

Bayard, D.; Weatherall, A.; Gallois, C.; and Pittam, J. 2001. Pax Americana? Accent attitudinal evaluations in New Zealand, Australia and America. *Journal of Sociolinguistics* 5(1):22–49.

BioWare. 2003. Star Wars: Knights of the Old Republic.

BioWare. 2009. Dragon Age: Origins.

BioWare. 2014. Dragon Age: Inquisition.

Brice, M. 2011. Speaking in Accents and the American Ethnocentrism in Video Games. http://www.popmatters.com/post/151275-speaking-in-accents-and-the-american-ethnocentrism-in-video-games/.

Costa, Y. M.; Oliveira, L. S.; and Silla, C. N. 2017. An evaluation of convolutional neural networks for music classification using spectrograms. *Applied Soft Computing* 52:28–38.

Ensslin, A. 2010. Black and white: Language ideologies in computer game discourse. *Language ideologies and media discourse: Texts, practices, politics* 205–222.

Ensslin, A. 2011. Recallin' Fagin: Linguistic accents, intertextuality and othering in narrative offline and online video games,. *Online Gaming in Context: The Social and Cultural Significance of Online Games* 56:224–235.

Espi, M.; Fujimoto, M.; Kinoshita, K.; and Nakatani, T. 2015. Exploiting spectro-temporal locality in deep learning based acoustic event detection. *EURASIP Journal on Audio, Speech, and Music Processing* 2015(1):26.

Graves, A.; Mohamed, A.-R.; and Hinton, G. 2013. Speech recognition with deep recurrent neural networks. In *Proceedings of 2013 IEEE International Conference on Acoustics, speech and signal processing (ICASSP)*, 6645–6649.

Hautamäki, V.; Siniscalchi, S. M.; Behravan, H.; Salerno, V. M.; and Kukanov, I. 2015. Boosting universal speech attributes classification with deep neural network for foreign accent characterization. In *Proceedings of Sixteenth Annual Conference of the International Speech Communication Association*.

Huang, Z.; Dong, M.; Mao, Q.; and Zhan, Y. 2014. Speech emotion recognition using cnn. In *Proceedings of the 22nd ACM international conference on Multimedia*, 801–804.

id Software and Grey Matter Interactive. 2001. Return to Castle Wolfenstein.

Irvine, J. T., and Gal, S. 2009. Language ideology and linguistic differentiation. *Linguistic anthropology: A reader* 402–434.

KingsIsle Entertainment. 2008. Wizard 101.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Lionhead Studios. 2004. Fable.

Lionhead Studios. 2005. Black & White 2.

Lippi-Green, R. 2012. *English with an accent: Language, ideology, and discrimination in the United States*. Routledge.

NCsoft. 2008. Aion.

Poos, D. 2013. Why is The Villain Always British? *GRIN*.

Schniedewind, N., and Davidson, E. 2000. Linguicism. *Readings for social diversity and social justice: An anthology on racism, anti-Semitism, sexism, heterosexism, ableism, and classism* 129–130.

Weinberger, S. 2015. Speech Accent Archive. http://accent.gmu.edu/.