Educational Administration: Theory and Practice
2024,30(4), 1455-1461
ISSN:2148-2403
https://kuey.net/

Research Article

A Review Paper On Developing A Real-Time Deepfake Voice Synthesis Framework: A Study In Artificial Intelligence

Chandrapal Singh Arya[1*], Dr. Ritu Sindhu[2]

[1*]Research Scholar, Department of Computer Science & Engineering, Lingaya's Vidyapeeth (A Deemed-to-be University), aryanchandrapal@gmail.com
[2]Professor, Department of Computer Science & Engineering, Lingaya's Vidyapeeth (A Deemed-to-be University), dr.ritusindhu@lingayasvidyapeeth.edu.in

Citation: Chandrapal Singh Arya et.al (2024). A Review Paper On Developing A Real-Time Deepfake Voice Synthesis Framework: A Study In Artificial Intelligence, *Educational Administration: Theory and Practice*, *30*(4), 1455-1461, Doi: 10.53555/kuey.v30i4.1692

ARTICLE INFO

ABSTRACT

Artificial intelligence, particularly Machine Learning and Deep Learning techniques, are becoming more prevalent in today's technological and social landscape. These advances have significantly aided the development of Speech Synthesis, also known as Text-To-Speech, in which speech is artificially produced from text using computer technology. This is where Voice Cloning technology comes in, which allows for the creation of artificial synthetic speech that resembles a specific human voice.
The study will start by analysing how several deep learning methods, such as recurre nt neural networks (RNN) and transformers can be used for voice synthesis. These methods' efficacy will be contrasted with that of more conventional voice synthesis te chniques.The project will thereafter concentrate on developing and putting into use a realtime deepfake speech synthesis system.The synthesis quality and realism of the suggested framework will be assessed. To boost the effectiveness of the suggested fr amework, transfer learning will also be investigated. The planned study will evaluate the suggested framework's possible uses in the media and entertainment sectors. Th e possibility for harmful or dishonest use of the technology will also be investigated, along with the ethical and societal ramifications of the suggested framework. Additio nally, the research will investigate the usage of GANs and multispeaker models for v oice synthesis in deepfake technologies. The proposed framework's scalability and ca pacity for handling huge datasets will be investigated. Finally, the research will investigate the use of deepfake framework to improve accessibility for individuals with speech impairments. Overall, this system aims to make significant contributions to the field of deepfake voice synthesis, and to provide a framework for the real time generation of high quality and realistic synthetic voices.
Arealtime deepfake speech synthesis framework will be investigated and developed a s part of this research project employing
Artificial intelligence. Industry sectors including entertainment, media, and custome r service could be completely transformed by the adoption of deep learning techniqu es for speech synthesis in deepfake technology. To accomplish highquality and realis tic speech synthesis in realtime, however, several issues still need to be resolved.

Keywords— Voice Cloning, Audio Deepfakes (ADs), Machine Learning (ML), Deep Learning (DL), Artificial Intelligence (AI), Speech Synthesizer, TTS.

I. INTRODUCTION

By using artificial intelligence techniques to build a machine learning speech model on genuine recordings, voice cloning produces a voice that is very similar to the original. This technology enables you to impersonate someone else, as was the case with a British energy firm where an unidentified hacker group employed artificial intelligence voice cloning technology to place phone calls and was successful in stealing 220,000 euros. For this voice must be:

- Crispness of sound
- Accurate language comprehension and pronunciation
- Reliable language sources

In order to create synthetic speech audio that accurately mimics the sound of a real person's voice, a real-time deepfake voice synthesis framework would need to use artificial intelligence methods like deep learning. There may be uses for this in the media, entertainment, and communication technology sectors.
Creating such a system would require amassing a sizable dataset of audio recordings of genuine human voices, which could then be used to train a deep learning algorithm. Using a WaveNet, a deep neural network capable of producing high-quality audio output that sounds quite natural, is one preferred method for this kind of activity.
To avoid the technology being used for evil objectives like impersonation, fraud, or misconduct, it is crucial to include strong quality control and verification procedures while creating a real-time deepfake voice synthesis framework.
When creating AI systems, ethical considerations and knowledge of the potential hazards and difficulties that could outweigh the seemingly limitless rewards are essential.

II. RESEACH OBJECTIVES

2.1: Blueprint

The objective of developing a real-time deepfake voice synthesis framework is to apply artificial intelligence and machine learning techniques to create high-quality synthetic voice for various applications such as speech synthesis, text-to-speech conversion, and voice cloning. The framework will use deep learning models to analyze and replicate the unique characteristics of real human voices in real-time. The project aims to advance the state-of-the-art in voice synthesis technology and enable new use cases for synthetic voice applications. The study will involve training deep learning models on large datasets of human speech and evaluating the performance of the resulting models in terms of quality and accuracy. Other objectives may include developing efficient algorithms for real-time processing and optimization techniques for reducing computational requirements.

2.2: To provide an overview of the review, the following research question was addressed:

The research questions that could be addressed in Developing a Real-Time Deepfake Voice Synthesis Framework: A Study in Artificial Intelligence could include:
RQ1: How can deep learning models be used to analyze and replicate the unique characteristics of human voices in real-time?
RQ2: How can the quality and accuracy of synthetic voice generated by the framework be evaluated and improved?
RQ3: What are the ethical and moral implications of implementing deepfake voice synthesis technology?
RQ4: What are the potential applications for real-time deepfake voice synthesis, and how can the technology be optimized for those applications?
RQ5: What computational requirements are needed for real-time deepfake voice synthesis, and how can algorithms be optimized to reduce computational cost?

2.3: Research Objectives:

1: To investigate and evaluate the use of various learning techniques for real-time voice and text-to-speech (TTS) synthesis in deepfake technology.
2: To access the potential applications of the proposed framework in industries such as entertainment, education, media, and customer service.
3: To improve the deepfake framework system accessibility for individuals with speech impairments.
4: To investigate the use of transfer learning for improving the performance of the deepfake framework.
5: To design and implement a text-to-speech (TTS) synthesis framework using neural networks.

2.4: Retrieving and selecting apropos literature:

The technological and social world of today is becoming more and more populated with artificial intelligence, particularly with machine learning and deep learning approaches. These developments have greatly influenced the creation of Speech Synthesis, often known as Text-To-Speech, which uses computer technology to artificially synthesize speech from text. This is where voice cloning technology comes into play, allowing for the creation of synthetic artificial speech that closely matches a chosen human voice. The quality of synthetic speech is currently improving thanks to Deep Learning and Artificial Intelligence (AI) developments. TTS applications are increasingly often submitted. Anyone who has used an IVR system for a phone, Apple's Siri, Amazon Alexa, auto navigation systems, or any of the numerous other voice interfaces has encountered synthetic speech. Their two methods of TTS in the past.
The first, called Concatenative TTS, compiles a library of words and sound units (phonemes) from audio recordings that may be combined to make sentences. It lacks the inflection and emotion that characterize authentic human speech. When employing Concatenative TTS, it takes a significant amount of work to clone any specific voice using this technique.
The second method, known as Parametric TTS, uses statistical models of speech to make the process of synthesizing a voice simpler and less expensive than Concatenation.

III. LITERATURE REVIEW

1. Neural Voice Cloning with a Few Samples: In this article, we present a neural voice cloning system that only requires a small number of audio samples as input. Both speaker encoding and speaker adaptability are topics of our work. A few clone examples are used to fine-tune a multi-speaker generative model for speaker adaption. completed by Sercan O. Ark * 1 Yanqi Zhou 1 Jitong Chen 1 Kainan Peng 1 Wei Ping.
2. Deepfakes Generation and Detection: Current State, Open Challenges, Solutions, and Future Directions: gives a thorough overview and in-depth examination of the machine learning (ML) based tools and deep fake generating technologies now in use, as well as the methods for identifying such manipulations for both audio and visual deep fakes.
We go through details about manipulation techniques, current publicly available datasets, and important benchmarks for measuring the effectiveness of deepfake detection techniques along with their findings for each type of deepfake. Aun Irtaza5, Momina Masood1, Mariam Nawaz2, Khalid Mahmood Malik3, Ali Javed4,
3. DATA EFFICIENT VOICE CLONING FOR NEURAL SINGING SYNTHESIS: We modify one of these methods for singing synthesis. Small quantities of target data can effectively adjust the model to brand-new, unheard voices by first using data from multiple speakers to build a multi speaker model. Ryunosuke Daido, Jordi Bonada, and Merlijn Blaauw completed the work.
4. Multilingual Speech Synthesis and Cross-Language Voice Cloning for Learning to Speak Fluently in a Foreign Language We offer a Tacotron-based multi speaker, multilingual text-to-speech (TTS) synthesis model that can generate high-quality voice in numerous languages. Additionally, the model can synthesized fluent Spanish speech using the voice of an English speaker without having been trained on any bilingual or parallel samples. Such transfer operates between languages that are not even remotely related, like English and Mandarin. Ye Jia, Andrew Rosenberg, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, and Bhuvana Ramabhadran completed the work.
5. Combining Unit-Selection and Statistical Parametric Speech Synthesis for Automatic Voice Cloning: In this presentation, we will describe two cutting-edge systems: the commercial unit-selection system Cere Voice, produced by Cereproc, and the HMM-based system HTS-2007, developed by CSTR and Nagoya Institute Technology. Using publicly available audio from the web, these systems have been used to replicate the voice of George W. Bush, the 43rd president of the United States.

IV. VOICE CLONING TECHNOLOGY

Voice cloning technology allows the generation of audio that imitates or mimics the voice of a real human being. There are different types of voice cloning technology, including technologies that require a human to record audio samples, and those that use text-to-speech (TTS) algorithms powered by artificial intelligence (AI) to generate synthetic audio.
Voice cloning technology has been used in various fields, such as in film and entertainment, where voice actors and impressions can be used to dub movies, TV shows, or video games. Additionally, voice cloning technology can be used to create personalized virtual assistants or chatbots that sound like real people.
There have also been instances where voice cloning technology has been used for malicious purposes, such as impersonation or fraud. The Federal Trade Commission has hosted workshops to examine voice cloning technologies and their potential impact on consumers.
Overall , voice cloning technology is a rapidly developing field that has the potential to revolutionize the way we interact with audio and virtual assistants, but its development and use need to be carefully monitored to prevent misuse.

A. Artificial Intelligence, Machine Learning, and Deep Learning

Artificial intelligence (AI), Machine Learning (ML), and Deep Learning (DL) are related terms but refer to different concepts. AI can be thought of as a broad field that encompasses any technique or algorithm that enables computers to perform tasks that typically require human intelligence, such as visual perception, speech recognition, decision-making, and language translation.
Machine learning is a subset of AI that involves training a computer to learn from data, without being explicitly programmed. This is often accomplished using algorithms that can identify patterns and make predictions based on patterns in the data.
Deep learning is a subset of machine learning that uses artificial neural networks, which are composed of layers of interconnected nodes, to learn and make predictions. Deep learning algorithms are particularly effective for tasks that depend on large amounts of data, such as image and speech recognition.

Identify applicable funding agency here. If none, delete this text box.

B. Batch Size

Neural networks are trained using stochastic gradient descent optimisation.[14] Based on the model's current state, a prediction is created, and the difference between the forecast and the expected values is used to estimate the error gradient. This error gradient is then used to update the model weights, and the procedure is repeated. The error gradient is a statistical estimate. The more training instances utilised in the estimate, the more accurate it is and the more probable it is that the network's weights will be altered in a way that improves the model's performance. The improved estimate of the error gradient comes at the expense of producing numerous extra estimates using the model.

"Minibatch gradient descent" refers to batches of a size that is between more than one example and the quantity of examples in the training dataset. The training dataset's batch size is set to more than one and below the total number of examples. Batch size figure in [15] [14]

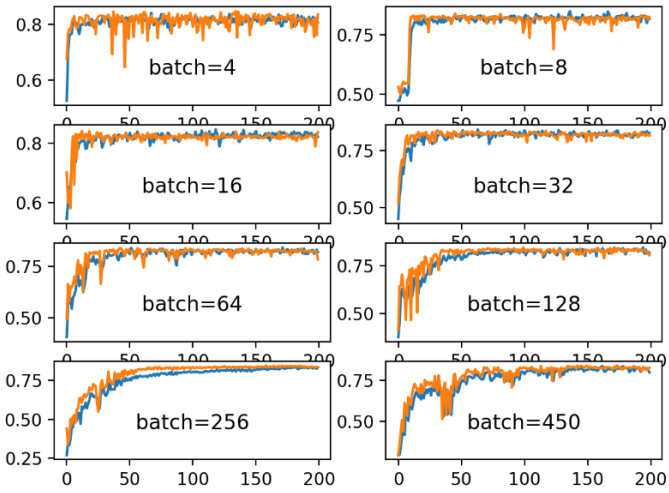

Fig. 2 Batch Size

C. Principles of Voice Cloning operation

The speech synthesis language sources should be determined. The process of turning text into synthetic speech that closely resembles real speech in accordance with the pronunciation standards of a special language is known as text to speech synthesis. Text to speech (TTS) systems are what these devices are known as. Text serves as the system's input element, and synthetic voice serves as its output element. There are two potential scenarios. The required speech material is simply pre-recorded when it comes to pronouncing the small number of phrases (and their linear pronouncing does not vary.

In this situation, specific issues are created. For instance, this method does not allow for the text to be read aloud when the text is unknown. The spoken text must be stored in computer memory for this function. Additionally, it will result in an increase in the amount of memory needed to store information. In the case of a lot of information, this will result in an essential load on the computer's memory and may cause certain operational issues. The primary strategy utilised in this paper is the voicing of previously unidentified text using a particular algorithm.

Each language has distinctive qualities of its own. For instance, the English language contains some inconsistencies between

particular letters and sounds. As a result, when two different letters are combined, they sound different from when they are used alone. For instance, the sound of the letters (t) and (h) independently differs from that of the chain (th). This is just one issue with the English language. In other words, the position of the letters determines whether or not they should be uttered. Thus, the first letter (k) of the word (know) is not spoken in accordance with the phonetic standards of the English language. Additionally, there are certain pronunciation characteristics in Russian. The letter (o) does not always pronounce similar sound, it should be indicated up front.

The principles of voice cloning operation utilize speech synthesis technology to create a synthetic voice that mimics the unique characteristics of a real human voice. This includes analysing and replicating aspects such as pitch, tone, cadence, and accent. The process of voice cloning involves training machine learning models on large datasets of human speech in order to recognize patterns and generate synthetic speech that sounds natural and intelligible. The resulting synthetic voice can be used for a variety of purposes, including text-to-speech conversion, speech synthesis, and voice cloning. Voice cloning technology has many potential applications, including assisting people with speech disorders, creating personalized voice assistants, and enhancing the entertainment industry by allowing actors and musicians to create digital versions of themselves. However, the development of voice cloning technology also raises ethical concerns, particularly in terms of privacy and the potential for misuse. As a result, it is important for researchers and developers to consider these implications and work to develop responsible solutions that prioritize the safety and autonomy

of individuals. Therefore, the principles of voice cloning operation involve utilizing advanced machine learning algorithms to create high-quality synthetic voice that replicates the unique characteristics of real human speech. While there are both potential benefits and ethical concerns associated with this technology, continued research and development holds promise for improving the quality of synthetic voice and expanding its potential applications.

D. Possible Token type Text

The set of possible token types for raw text in a table will vary depending on the specific context and purpose of the table. However, some common token types that may appear in text tables include:
Words: individual units of text that are separated by whitespace, punctuation, or other delimiters.
Numbers: numeric values that may be integers, decimals, or scientific notation.
Dates and times: values that represent specific points or ranges of time.
Symbols: non-alphanumeric characters that have specific meanings, such as currency symbols or mathematical operators etc. shown in the table below:

Type	Text	Speech
Decimal numbers	1.2	One and Two Tenth
Ordinal Numbers	1-st, 2-nd	First, Second
Roman numbers	III, X	Third, Tenth
Alphanumeric Strings	1^{10}	One a Power of Ten
Phone Numbers	+918454968480	Plus, Nine, One, Eight, Four, Five, Four, Nine, Six, Eight, Four, Eight, Zero
Count	99	Ninety Nine
Date	21/10/2023	Twenty First of October Twenty Twenty Three
Time	11:15 pm	Quarter past eleven post meridiem
Mathematical	2+8=10	Two plus eight is equal to Ten

Table. 1: Possible Tokens

E. Speech versus text difference

The main difference between speech and text lies in their modality. Speech is a mode of communication that involves the use of spoken language and sound waves that are transmitted through the air or other medium. Text, on the other hand, is a written mode of communication that represents language through symbols such as letters, numbers, and punctuation marks.
Speech and text signals both have a distinct hierarchical structure. We can infer from hierarchical representation that a model of the mechanism of speech generation must be created in order to create systems of speech synthesis that are qualitatively constructed. The information flow within the system should be initially defined and should follow the plan shown below:

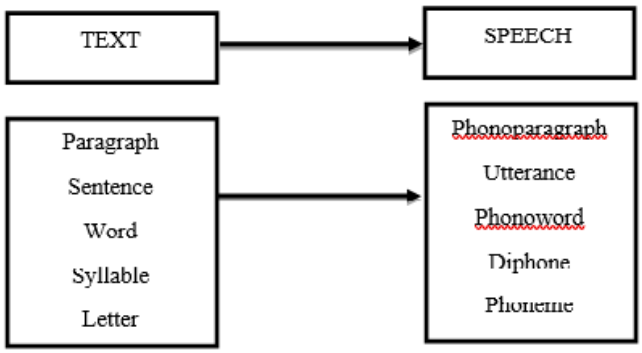

Fig. 2: Text vs Speech

F. Survey of Current Apps and Websites

Based on the search results, here are a few current apps and websites that utilize voice cloning technology:
Parodist- an app with the voices of over 40 famous artists, personalities, and cartoon characters. It allows users to enter the first and last name of a friend or family member and hear them speak in a familiar voice.
Amazon Polly - a service that uses deep learning technologies to synthesize natural-sounding human speech from text. It can be used to convert articles or other written content into audio formats.
Real time voice changer - a free app for PC that allows users to modify their voice in real-time or clone any voice they want some news brands, such as BBC, have started using voice cloning technology to create synthetic voices for their news broadcasts and podcasts. There are some samples are:

Web	n. characters	n. voice	language	Time	download	Free	speed
ttsmp3	3,000	61	28	no	Yes	yes	no
fromtexttospeech	50,000	17	8	yes	Yes	yes	yes
ibm	5,000	40	13	yes	no	yes	yes
fakeyou	1000	1385	5	yes	yes	yes	yes
IOS/ANDROID	n. characters	n. voice	language	Time	download	Free	speed
MOTOREAD	2500	1	9	No	no	Yes/no	No
Voice dream reader	5000	15	11	No	No	no	Yes
Voice aloud reader	5500	2	15	Yes	No	yes	No
Speech central	3000	2	5	No	No	Yes/no	yes

Table. 2: Survey Data

G. *The Software's Libraries Used*

Libraries	Description
Tkinter	is the standard GUI library for Python, when combined with Tkinter provides a fast and easy way to create GUI applications.
Pyttsx	is a text-to-speech conversion library in Python, it is a very easy to use tool which converts the entered text into speech.
Winsound	module provides access to the basic sound-playing machinery provided by Windows platforms
Sounddevice	module provides bindings for the Port Audio library and a few convenience functions to play and record NumPy arrays containing audio signals.
Scipy	It provides more utility functions for optimization, stats and signal processing.
Wavio	reads a WAV file and returns an object that holds the sampling rate, sample width (in bytes), and a numpy array containing the data. write writes a numpy array to a WAV file, optionally using a specified sample width.
Pygame	As pygame and the SDL library are portable across different platforms and devices, they both need to define and work with abstractions for various hardware realities. Understanding those concepts and abstractions will help you design and develop your own
Numpy	is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, Fourier transform, and matrices
Matplotlib	is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy.
Ipython	provides a rich architecture for interactive computing with: A powerful interactive shell. A kernel for Jupyter. Support for interactive data visualization and use of GUI toolkits.
Pil	a free and open-source additional library for the Python programming language that adds support for opening, manipulating, and saving many different images file formats.

Table. 3: List of Libraries

V. RESULTS AND OUTCOMES

There are a few mentions of deepfake voice synthesis frameworks and AI in the search results, but no specific study or outcomes mentioned in regards to a real-time deepfake voice synthesis framework. One search result mentions a "new wave of deepfake detection research directed against real-time facial synthesis systems", but this is focused on visual deepfakes rather than voice . Overall, the search results suggest that voice cloning and deepfake technology are actively being developed and explored in the realm of AI and machine learning, but there is a growing awareness of the ethical implications and concerns about misuse.

VI. CONCLUSION AND RECOMENDATIONS

As an AI language model, I cannot provide an opinion or analysis on specific research articles. However, I can explain the general approach and significance of the research you mentioned.
In the context of AI research, developing a real-time deepfake voice synthesis framework is a significant achievement. This research aims to generate artificial speech that sounds like a specific person by using a deep learning algorithm. It involves training the model on the target speaker's voice to generate a synthetic voice that mimics their vocal characteristics.
The potential applications of such a technology are vast, from creating voiceovers for movies and commercials to personalizing virtual assistants and speech-based communication. However, there are also ethical and privacy concerns that must be considered.
Based on the research, recommendations for further study might include improving the accuracy and naturalness of the synthetic voice, developing better methods for detecting deepfake voices, and exploring ways to mitigate the potential negative impacts of this technology.
Therefore, the development of a real-time deepfake voice synthesis framework is a significant advancement in the field of AI, with both potential benefits and potential risks that should be carefully considered.

VII. REFERENCES

1. Expressive Neural Voice Cloning. Proceedings of Machine Learning Research 157:–, 2021, Paarth Neekhara* Shehzeen Hussain* Shlomo Dubnov ,Farinaz Koushanfar, Julian McAuley ,University of California, San Diego, 9500 Gilman Dr, La Jolla, CA 92093 * Denotes Equal Contribution

2. Combining Statistical Parameteric Speech Synthesis and Unit-Selection for Automatic Voice Cloning by Matthew P. Aylett12, Junichi Yamagishi1, 1Centre for Speech Technology Research, University of Edinburgh, U.K.2Cereproc Ltd., U.K.
3. TTS-SYNTHESIZER AS A COMPUTER MEANS FOR PERSONAL VOICE “CLONING” Boris M. Lobanov* and Helena B. Karnevskaya** Institute of Engineering Cybernetics, Nat. Ac. of Sc. Belarus * Minsk Linguistic State University
4. Text Analysis and Word Pronunciation in Text-to-speech Synthesis, Mark Y. Liberman, Kenneth W. Church, AT&T Bell Laboratories, 600 Mountain Ave. Murray Hill, N.J., 07974
5. The Main Principles of Text-to-Speech Synthesis System by K.R. Aida–Zade, C. Ardil and A.M. Sharifova
6. RCHISEGMENT-BASED LETTER-TO-PHONE CONVERSION FOR CONCATENATIVE SPEECH SYNTHESIS IN PORTUGUESE, Eleonora Cavalcante Albano and Agnaldo Antonio Moreira, LAFAPE-IEL-UNICAMP, Campinas, SP, Brazil
7. Sadhan ¯ a¯ Vol. 36, Part 5, October 2011, pp. 837–852. c Indian Academy of Sciences, An introduction to statistical. parametric speech synthesis SIMON KING, The Centre for Speech Technology Research, University of Edinburgh, Edinburgh
8. DATA EFFICIENT VOICE CLONING FOR NEURAL SINGING SYNTHESIS by Merlijn Blaauw, Jordi Bonada, and Ryunosuke Daido, Music Technology Group, Universitat
9. Pompeu Fabra, Barcelona, Spain, Sound Processing Group, Yamaha Corporation, Hamamatsu, Japan
10. (SV2MTTS) Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis by Ye Jia∗ Yu Zhang∗ Ron J. Weiss∗ Quan Wang Jonathan Shen Fei Ren Zhifeng Chen Patrick Nguyen Ruoming Pang Ignacio Lopez Moreno Yonghui Wu Google Inc. {jiaye,ngyuzh,ronw}@google.com
11. Efficient Neural Audio Synthesis by Nal Kalchbrenner*1Erich Elsen * 2 Karen Simonyan 1 Seb Noury 1 NormanCasagrande 1 Edward Lockhart 1 Florian Stimberg 1 Aaron van den Oord ¨ 1 Sander Dieleman 1 Koray Kavukcuoglu
12. TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS by Yuxuan Wang∗, RJ Skerry-Ryan∗ , Daisy Stanton, Yonghui Wu, Ron J. Weiss† , Navdeep Jaitly, Zongheng Yang, Ying Xiao∗ , Zhifeng Chen, Samy Bengio† , Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous∗ Google, Inc. {yxwang,rjryan,rif}@google.com
13. GENERALIZED END-TO-END LOSS FOR SPEAKER VERIFICATION Li Wan Quan Wang Alan Papir Ignacio Lopez Moreno Google Inc., USA {liwan, quanw, papir, elnota} @google.com
14. Control the Stability of Training Neural Networks with the Batch Size © 2022 Machine Learning Mastery. All Rights Reserved.
15. Manning Publications Co. Development editor: Toni Arritola 20 Baldwin Road Technical development editor: Jerry Gaines PO Box 761 Review editor: Aleksandar Dragosavljevic´ Shelter Island, NY 11964 Printed in the United States of America 1 2 3 4 5 6 7 8 9 10 – EBM – 22 21 20 19 18 17