Propagation of Speech Sounds in SPREAD

Jiali Sheng

Advisor: Norman Badler, PengFei Huang, Mubbasir Kapadia

University of Pennsylvania

ABSTRACT

SPREAD is a novel agent-based sound perception model that was just presented in the paper "SPREAD: Sound Propagation and Perception for Autonomous Agents in Dynamic Environments". Although SPREAD tests the 100 different environmental sounds, it has never been tested on speech sounds. Speech sounds are important in a perception system to support character animation and crowd simulation because the majority of human interaction happens when we are speaking to each other. The range of possible signals for speech sounds are much smaller, as a result, Human listeners are much more sensitive to the tiny nuisances. This project seeks to extend SPREAD to detect such nuisance by finding and passing along more useful acoustic signals to the propagation step and creating a larger database for the recognition step.

Project Blog: http://jollysound.blogspot.com/

1. INTRODUCTION

Imagine a large party where a floor of people are conversing all at once. All the sudden, you hear someone say your name from behind you so you turn to face the speaker. This describes the selective attention effect (also known as the cocktail party effect). It is what allows people to detect important events in an unattended stimuli. It is also an important part of human behavior.

In a system that tries to animate behavior based on each agent's perceived auditory sense, it is not enough to simply model environment sounds. So much of human behavior deals with spoken language that speech sounds should naturally be the next step of expansion. With support for speech sounds, the system can then simulate important behavioral phenomenons such as the cocktail party effect.

Speech sounds are so much more difficult to differentiate between than environmental sounds. In linguistics, the smallest unit of categorization is the phoneme. Unlike environment sounds where it may be enough to distinguish only between impact sounds or friction sounds, speech sounds require a more delicate set of categorization. For example: the duration of a phoneme, the frequency ratio of the different formats in the signal, and the various versions of a phoneme when it is next to various other phonemes.

Problem Statement. What needs to be done so that SPREAD can propagate and perceive speech sounds.

Motivation. In the real world, one of the most forms of communication is speech sounds. Through talking and communicating, one person can trigger a different set of behaviors in another person. SPREAD is a system that tries to use sound propagation as a means of animating agents in a virtual environment. With the ability to propagate and perceive speech sounds, we are creating the opportunity to simulate a new and bigger set of behaviors in virtual agents.

Proposed Solution. Relevant frequency range for speech sounds range from 300 hz to 8,000 hz. This is a bigger range than what SPREAD currently support. As a result, first a way to scale the raw acoustic signal is needed. After scaling the raw acoustic signal, extraction of important frequencies needs to be more fine tune. Currently each SPREAD packet deals with only a select few frequencies in a small range. Determining speech sounds relies on the information regarding the ratio of formant. As a result, a method of detecting multiple packets per time sample and sending the multiple packets at the same time will be developed. In the receiving end, a solution for perception will be created based on recognition of each phoneme. The program will recognize phonemes based on either a HCA tree created with a phoneme matrix evaluated with the propagation step, or with a known speech transcriber.

Contributions.

This project makes the following contributions:

- Finds a way to represent speech sounds based on SPREAD packets (with none or very little alteration to the existing model)
- Research and implement the best way for the agent to perceive the propagated packets.
- Develop a working pipeline for speech perception in SPREAD

1.1 Design Goals

The purpose of the project is to research a working model to speech sound propagation. It currently targets SPREAD, but the concept could also be used in other systems.

1.2 Projects Proposed Features and Functionality

The project will result in an additional step in the existing pipeline during packet generation phase such that in the event of a speech sound, SPREAD will generate a different set of packets than if the event is an environmental sound. In the perception step, SPREAD will use a different piece of the pipeline when it receives packet representing a speech sound.

2. RELATED WORK

Some prior work in terms of making speech sounds more robust to degradation in distance exists in the field of telephone networks.

“**Sources of Degradation of Speech Recognition in the Telephone Network” by Pedro J, Moreno and Richard M. Stern** is a study done to categorize how speech degrades over the telephone network. There are many other such related works in speech recognition related to telephone networks done I nthe 80s and eaerly 90s. Another example is “**Phonetic Classification on Wide-Band and Telephone Quality Speech” Sources of Degradation of Speech Recognition in the Telephone Network”** by Chigier, B.

2.1 Reference Material

1. **“Music and Computer” by Douglas Repetto, Polansky, burk, Roberts, Rockmore.**

   http://music.columbia.edu/cmc/musicandcom puters/

This is basically a textbook in web format. It serves as a wonderful introduction about how sound is represented digitally. It briefly introduces sampling, the frequency domain, synthesis, and sound editing. The best thing about this website is the fact that each section has a widget that the reader can play with and get a more intuitive understanding of the material in the section.

2. **“The Voice in the Machine: Building Computers That Understand Speech” by Roberto Pieraccini**

Some information on speech recognition, and an overview on all the research that has been done so far.

3. **“Audio Anecdotes: Tools, Tips, and Techniques for Digital Audio” by Ke Greenebaum and Ronen Barzel**

A good reference tool for audio techniques.

4. **“Computational Methods in Acoustics” by Ulf R. Kristiansen and Erlend M. Viggen**

This is is an over view (more technical than the previous sources) on the techniques in sound propagation.

5. **“Interactive Physically-based Sound Simulation” by Nikunj Raghuvanshi**

A Thesis on sound simulation. There are some content on sound generation but most of it deals with propagation.

6. **“The Sounds of Language” by Henry Rogers**

A text book on how spoken language is categorized, analyzed, and studied.

3. PROJECT PROPOSAL

I propose a project that will be the start of a solution to speech propagation and recognition in SPREAD. This project will focus on American English, and use the 44 phonemes to find out few important traits that distinguish one phoneme from another. SPREAD will be fitted with two new parts in it's pipeline that is especially designed for the propagation and perception of speech sounds.

3.1 Anticipated Approach

The Approach can be divided into two parts: packet generation and perception.

In the packet generation step, SPREAD will generate multiple packets per time step to accommodate for the various formants needed in phoneme recognition. To do so, first, tests will be conducted by first putting each English phoneme through the original packet generator and seeing what information is missing from the packets after propagation. The algorithm will be tweaked to allow for multiple packets to be generated per time frame.

In the perception part of the project, propogated signals will first be put through a known speech transcriber and evaluate the effectiveness. It is very likely that an out-of-the-box voice speech transcriber will not be calibrated for SPREAD, so instead, the speech transcriber will generate a phoneme confusion matrix. With the phoneme confusion matrix, we will create an HCA tree and evaluate how SPREAD does in terms of phoneme recognition given the tree.

3.2 Target Platforms

This project will be using Unity and Mat lab on windows. Therefore targets Windows systems,

3.3 Evaluation Criteria

The evaluation criteria will be to animate a very simple version of the cocktail party effect. Each agent will be set to always look for a sequence of phonemes that represents their name. In the event that a speech sound representing an agent's name is propagated in the virtual environment, that agent is responsible to acknowledge the event.

4. RESEARCH TIMELINE

Project Milestone Report (Alpha Version)

⑤Initial tests in SPREAD

⑤Wrote some Matlab code that plays around with the tweaking of sound packets in the packet generation step.

⑤Have a vague idea of how to solve the problem. Or at least have started brainstorming ideas that would fix the problem.

Project Final Deliverable

List what you will deliver at the end of the semester

⑤Software for the extra parts in the pipeline that deals with speech sounds.

⑤Demo of Cocktail party Effect.

⑤Documentation of external library

Project Future Tasks

⑤This project targets the English language. It would be useful for the extra time to target other languages as well.