Propogation of Speech Sounds with SPREAD

Jiali Sheng

Advisors: Pengfei Huang, Mubbasir Kapadia, Norman Badler

Department of Computer and Information Science -- University of Pennsylvania

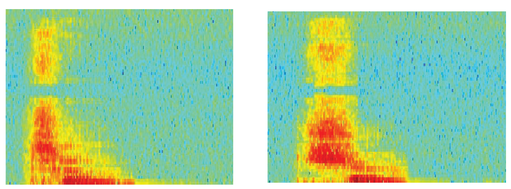
Senior Project Poster Day 2013

SPREAD

SPREAD is a novel agent-based sound perception model. It simulates how sound features are propogated, attenuated, spread and degraded as they traverse a virtual enviromenet.

Problem

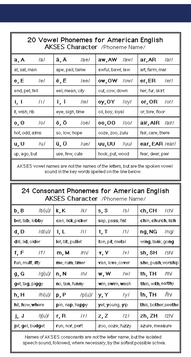
SPREAD has been demonstrated to work with 100 different enviroment sounds and can be the basis of agent-enviroment interactions. However enviroment sounds are not enough to simulate agent-to-agent or user-to-agent interactions.



Less differences between phonemes provide a bigger challange in the packet generation and matching stage (left: "t" ; right: "ch").

Goal: A user should be able call the name of an agent through speech, and the agent will either respond or not respond based on their location in the scene.

Phonemes-Approach



20 Vowel Phonemes for American English

AKSES Character	/Phoneme Name/				
a, A	/æ/	ā, A	/ae/	aw, AW	/aw/	ar, AR	/ar/
at, sat, mam		apa, pad, bame		awful, bawl, law		art, farm, star	
e, E	/e/	ē, E	/ee/	ew, EW	/ew/	er, ER	/er/
end, pet, fret		eel, meon, city		cut, cow, down		her, fur, stirt	
i, I	/i/	ī, I	/ie/	oy, OY	/oy/	or, OR	/or/
it, mim, rib		eye, sigh, time		oil, toy, boys		or, fore, floor	
o, O	/o/	ō, O	/oe/	oo, OO	/oo/	ar, AR	/ar/
hot, odd, ahms		so, low, hope		ooze, zoo, zuta		fur, cure, there	
u, U	/u/	ū, O	/ue/	uu, UU	/uu/	ear, EAR	/ear/
up, ugly, bull		ute, fine, cute		book, pull, wood		fear, dear, pier	

AKSES vowel names are not the names of the letters, but are the spoken vowel sound in the key words spelled on the line below.

24 Consonant Phonemes for American English

AKSES Character	/Phoneme Name/						
b, B	/b(u)/	k, K	/k/	s, S	/s/	ch, CH	/ch/
bet, bib, rabby		can, tick, cuker		sap, press, fat		chin, church, tich	
d, D	/d(u)/	l, L	/l/	t, T	/t/	ng, NG	/ng/
dit, bid, rider		let, tit, puller		tim, pit, metal		ring, tong, gong	
f, F	/f/	m, M	/m/	v, V	/v/	sh, SH	/sh/
fun, ruff, fifty		ate, rate, tater		van, cave, cover		she, crash, rovship	
g, G	/g(u)/	n, N	/n/	w, W	/w/	th, TH	/th/
get, tug, ziggy		ion, tick, funky		win, swim, wash		thin, with, mithy	
h, H	/h(u)/	p, P	/p(u)/	y, Y	/y/	th, TH	/th/
hit, how, where		pin, rap, happy		yet, young, yo		this, father, zoother	
j, J	/j(u)/	r, R	/r/	z, Z	/z/	zh, ZH	/zh/
jet, gel, badge		run, rot, part		zoo, ozzie, fuzzy		azure, measure	

Names of AKSES consonants are not the letter name, but the isolated speech sound, followed, where necessary, by the called possible schwa.

Spoken speech can be broken down. The smallest unit linguistically is the "phoneme. The English language consists of about 44 different phonemes.

Each agent is givent the shortest name possible -- one phoneme long. This makes the pipeline easy to analyze and debug.

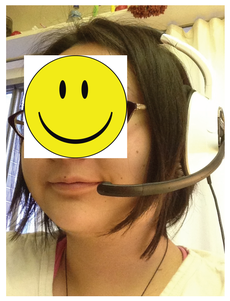
Pipeline

1. Speech **Input** from User
2. Program finds a matching phoneme from **database**, and propogate precomputed packet.

3. SPREAD propogates phoneme packet in the enviroment.
4. Sound Degrades as it travels through the enviroment and bounces off obstacles

5. Agent recieves packet
6. Agent tries to perform **Recognition**
7. If packet matches agent's name, agent responds with a behavior.

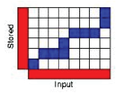
Input

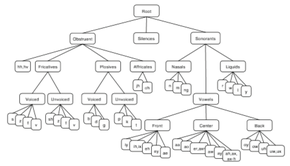


Speech Input from user is first processed through Speech SDK 5.1 for phonemes.

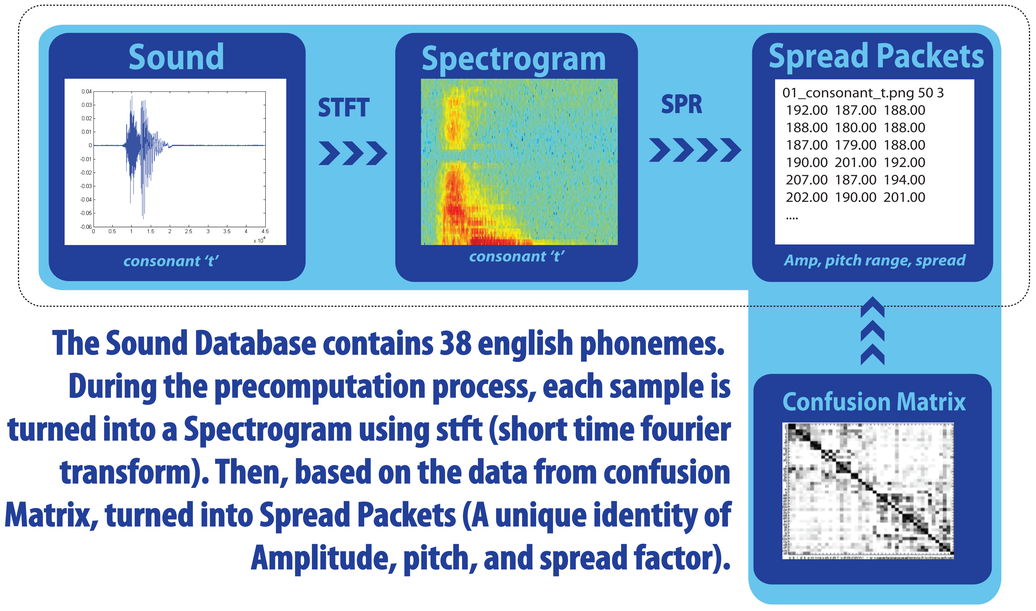
Recognition

Degraded Packets are first matched to original packets via **Dynamic Time Warping**. Confidence scores for matches are calculated using an **HCA tree** calculated from the Similarity Matrix.





Sound Database



The Sound Database contains 38 english phonemes. During the precomputation process, each sample is turned into a Spectrogram using stft (short time fourier transform). Then, based on the data from confusion Matrix, turned into Spread Packets (A unique identity of Amplitude, pitch, and spread factor).

Evaluation

Three agents with the names "t", "ah", and "m" are located in the scene. each agent was able to recognize their name when they are called. (In the screen shot, the propogated sound is "t". The agent responds by turning red).

However for some phonemes, analysis of second or third phoneme matches makes very little sense. (ie: when propogating "t", the second match will be the "ei" *eight)

Future Work

Sample using a logrithmic scheme (instead of a uniform scheme) to better simulate the biological perception model.

Implement multiple-phoneme propogation and recognition.

Introduce into the system, a statistics based guessing scheme that will predict the next phoneme in a sequence.