Speech Perception Using Real-Time Phoneme Detection: The BeBe System

Latanya Sweeney and Patrick Thompson

Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
sweeney@lcs.mit.edu, pmt@ai.mit.edu

Speech Perception Using Real-Time Phoneme Detection: The BeBe System

Latanya Sweeney and Patrick Thompson

Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
sweeney@lcs.mit.edu, pmt@ai.mit.edu

Abstract

We define a new approach to speech recognition based on auditory perception and modeled after the human brain's tendency to automatically categorize speech sounds [House 1962; Liberman 1957]. As background, today's speech recognition systems are knowledge-driven since they require the existence of word and syntax-level knowledge to identify a word from the sound. In contrast, our system uses no higher-level knowledge. Its architecture consists of competing parallel detectors which in real time identify phonemes in the waveform. Each detector, which is a simple algorithm, continuously samples the sound and reports the degree to which the samples contain its designated phoneme. The phoneme detector with the highest precedence and the greatest certainty above a minimal threshold prevails and its phoneme is added to an output queue. In preliminary experiments, four such detectors were tested and they properly identified 83-100% of their designated phonemes in both discrete and continuous speech, independent of the speaker, suggesting that an overall system which incorporates our approach would be much more robust and flexible than traditional systems.

Keywords: speech recognition, phoneme detection, speech enhancement, auditory perception, AI architectures, speech understanding, real-time systems.

Acknowledgments

We gratefully acknowledge Gerald Sussman for discussions, support and enthusiasm. We also thank Norman Margolus, Ken Yip, Ken Stevens, Robert Berwick, Gina Levow, Tom Knight and Patrick Winston for discussions; Eric Jordan and Philipp Schmid for editorial comments; and Peter Szolovits for providing an environment that made it possible for us to explore our own ideas. Henry Leitner inspired this work and we are grateful for his continued support and for that of Harvard University DCE. This work has been further supported in part by a Medical Informatics Training Grant (1 T15 LM07092) from the National Library of Medicine.

Contents

1 Introduction

The task of speech recognition is to map a digitally encoded signal to a string of words. Over the past 10 years speech recognition technology has advanced dramatically, evolving into 65,000-word vocabulary research systems capable of transcribing naturally spoken sentences on specific topics from any new talker [Bourlard, et al., 1996], but achieving human-like performance remains distant. Lippmann [1996] points out: (1) the error rates of machines are more than an order of magnitude greater than those for humans under the most ideal circumstances for the machines; (2) machine performance plummets much faster than humans when operating in noise and other degraded conditions; (3) humans exhibit much more powerful types of adaptation and incorporate newly learned words; and (4) humans rely on context much less than machines and can accurately recognize nonsense words, which are words that sound like English words but in reality have no meaning.

We agree with Lippmann that improvements at the low, acoustic-phonetic level must be achieved if machines are to equal human performance on real-world tasks. Our challenge is to model human auditory reasoning and behavior, so we consider our approach in this work as one of auditory perception, which models how the identities of sounds, including speech sounds, are learned and processed by a human listener.

1.1 The BeBe System

The specific aim of the BeBe System is to reliably and consistently detect phonemes in continuous speech. Computation can be performed in real time and recognition is robust enough to adapt to different speakers and changes in talking speed. Phonemes are the basic sound groups of a language and most languages have 50 or fewer phonemes, so explicitly identifying phonemes in the waveform offers an economic representation.

Phonemes are language-specific, so the identification of phonemes in languages other than American English requires building different phoneme detectors. Also, there is tremendous variation among sounds within a phoneme, so we consider phonemes to be abstract classifications and the goal of the BeBe System to be the continuous classification of waveforms. BeBe's activity concerns what is traditionally termed the *preprocessing stage* of speech recognition, in which the original sound waveform is converted to a phonemic digital representation. Depending on the speech recognition system, there may be three or more stages until a word or phrase is recognized where each stage often inherits inexact and noisy results from earlier stages. As a result, ambiguity in the overall system can magnify exponentially. Having a reliable phoneme-based detector in the preprocessing stage clearly improves speech recognition systems since it dramatically reduces uncertainty throughout the system and in its best case reduces the speech recognition problem to one of looking up phoneme sequences in a table.

2 Background

One of the issues impeding widespread implementation of large-vocabulary, continuous-speech systems is computational complexity [Rabiner and Juang, 1993]; therefore, in providing background for our approach, we will first discuss computational architectures. Then we will look at continuous speech and the acoustic-phonetic approach to recognition since research in these areas also views words as sequences of phonemes. Following that, we will shift our attention to human speech perception.

2.1 Computational architectures

In the 1970s, the Advanced Research Projects Agency (ARPA) of the United States Department of Defense conducted its Speech Understanding Research Project (ARPA SUR) which catapulted speech recognition research from speaker-dependent, small-wordlist recognition to the large-scale language model systems available today. Only one system met ARPA SUR's original mandate, Carnegie Mellon University's Harpy [Klatt, 1977], and the success of Harpy's statistical modeling techniques continues to have a profound effect as researchers seek to build larger statistical knowledge bases in an attempt to overcome problems and extend the performance of systems. Some attention has been given to recognizing phonemes directly from the waveform in the preprocessing stage, but little attention has focused on the use of parallelism in the preprocessing stage to improve speed and accuracy and to mitigate the overall system's computational complexity.

Our BeBe System utilizes numerous detection algorithms competing in parallel to label contiguous samples of sound as being particular phonemes. Each detection algorithm recognizes a specific phoneme though there may be more than one detection algorithm for a phoneme. For example, there is a single detection algorithm for the phoneme /R/ which is found in words like early, hurt and stir. There is a detection algorithm for the sound [p] as in bumper, spit and culprit, and another detection algorithm for [p^{h}] as in pit, pain and part, since these two allophones are different sounds of the same phoneme /p/. Our convention is to write phonemes using Arpabet letters (ARPA's phonetic alphabet that can be typed using a traditional computer keyboard) between slashes (/ /) and to write phonetic symbols between square brackets ([ ]).

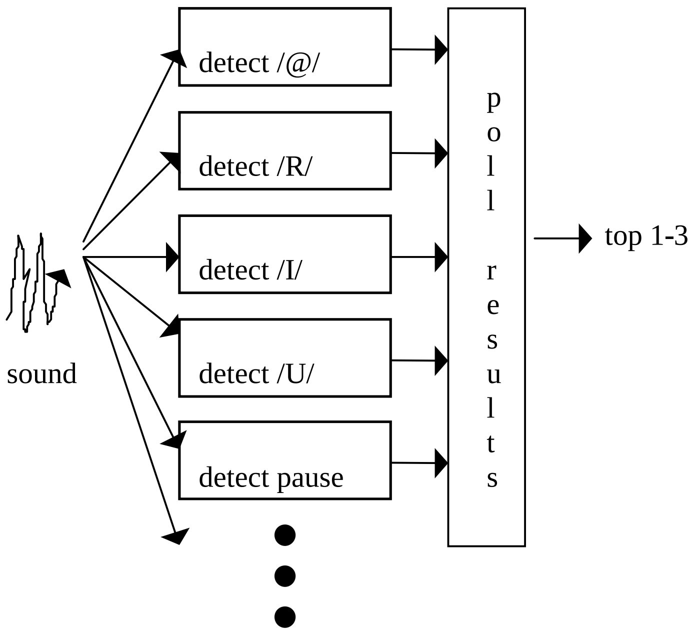

Diagram 1. Block diagram of BeBe System. The detection algorithms report the likelihood that the sound samples contain the designated phoneme. Examples of the sounds are: /@/ as in bat, /R/ as in stir, /I/ as in fit, and /U/ as in should.

Detection algorithms in BeBe share results and compete based on the certainty of their findings. Each algorithm tries to identify occurrences of its assigned phoneme, and reports for each sample in the waveform how likely it is that the sample is part of an instance of the algorithm's assigned phoneme. The algorithm with the highest precedence and the greatest certainty above a minimal threshold prevails and its results may be made available to all detection algorithms for future use. Diagram 1 presents an overview of the BeBe architecture which will be discussed in further detail in the Design and Implementation sections.

In the area of speech recognition, Hearsay-II's blackboard architecture [Erman, et al., 1980] engages multiple knowledge sources that work in parallel. Adjacent sources communicate with each other using a message center called a blackboard. This is similar to the BeBe System except communication is central to Hearsay-II because each level is believed to be so uncertain that a collaborative effort is required and not a competitive one.

Another approach similar to BeBe's is the Scrub System [Sweeney, 1996] which locates and replaces personally identifying information in medical records. Letters between physicians and notes written by clinicians often contain nicknames, phone numbers and references to other care-takers and family members, making it difficult to share medical records while still maintaining a commitment to patient confidentiality. The Scrub System used numerous algorithms competing in parallel to identify personal information in unrestricted text, and the system found 99-100% of these references. In contrast, the straightforward approach of global search-and-replace properly located no more than 30-60% of all such references. Although the Scrub System used numerous knowledge sources such as lists of area codes, first names, medical terms and so forth, BeBe has no stored lists and uses no higher-level predictive knowledge. BeBe relies only on its ability to recognize phonemes.

In Ether [Kornfeld, 1979], decentralized parallel processing was shown to be an effective alternative, in computational complexity terms, to many kinds of heuristic search strategies that implemented backtracking. Parallelism in Ether, as in BeBe, is design-based and does not necessarily require parallelism in its implementation. However, Ether does not use certainty factors as a scoring system between competing processes; instead the first process to complete the task solves the problem.

2.2 Continuous speech

In continuous speech, the speaker communicates in a natural manner with naturally occurring pauses. Most of the speech recognition systems commercially available today are really "connected speech" systems which require a deliberate pause between each word [Markowitz, 1996]. The pause cannot be eliminated since it is used to identify word boundaries. Table 1 shows a phonetic transcript of some American English phrases and their corresponding text with and without word boundaries. To combat this problem in continuous speech, many systems use triphones, a phoneme surrounded by contextual information on both sides, to model cross-word coarticulation. Researchers at AT&T Bell Laboratories [Pieraccini, et al., 1991] reported that when these highly detailed speech units were used the complexity of the overall implementation increased quadratically with the number of units, making a full-search implementation at that time "totally impractical, if not impossible."

Clearly, having a reliable phoneme detector changes the nature of speech recognition and reduces computational complexity. If a phoneme detector could provide a phonetic transcript that was 100% accurate in connected speech, then the remaining task for speech recognition would be to simply look up the phoneme sequence in a table which, like desktop dictionaries, associates phonetic pronunciation to spellings. As the phoneme detector becomes less reliable, however, the complexity of the remaining task increases. At, say, 90% accuracy, the task may be simple; at 80% accuracy, it may be reasonable; but, at 60% accuracy, recognition from the transcript would be difficult, depending of course on the nature of the inaccuracy. A similar computational complexity relationship holds for continuous speech as well. So, an approach to speech recognition we propose is to build a reliable phoneme detector.

	Phrase I
Phonetic Transcript	[gɛtθIkbʊk]
English text stream	getthickbook
English sentence	get thick book

	Phrase II
Phonetic Transcript	[s∂p^{h}owzp∂rejdk∂rowd]
English text stream	supposeparadecorrode
English sentence	suppose parade corrode

	Phrase III
Phonetic Transcript	[sp^{h}owzpr$_{1}$ejdkr$_{1}$owd]
English text stream	supposeparadecorrode supposeprayedcrowed
English sentence	suppose prayed crowed, suppose parade corrode

Table 1. A phonetic transcript of English phrases and their corresponding text with and without word boundaries. The last two phrases are the same utterances except phrase II is at normal speed and III is spoken rapidly.

BeBe's processing is a left-to-right, one-pass system. The entire waveform does not need to be saved. As a result, many more word boundaries are explicitly detected than is possible using traditional recognizers. In fact, one of the detection algorithms in BeBe is itself a pause detector reporting how likely it is that the sound sample is a pause. Of course the pause detector in BeBe still will not distinguish all word boundaries. These word divisions must be determined from the phonetic transcript BeBe produces. In the discussion section, we will present strategies that combat this, but first we will examine other phoneme-centered systems, human speech perception, and BeBe's implementation and run-time results.

2.3 Phoneme models

Acoustic-phonetic recognition involves global reasoning about the identity of phonemes in a digitized representation of a spectrogram [Cole, et al., 1980; Zue, 1985]. There are three stages: feature extraction, segmentation and labeling, and word-level recognition. First, the system examines the signal representation for features that describe spectral patterns. Extracted features are then interpreted using acoustic-phonetic rules that attempt to label the phoneme and segment where the phoneme begins and ends, but this is often very uncertain since the feature set does not distinguish well enough between similarities in phonemes and coarticulation effects to be reliable. The uncertainty in the results leads to a set of hypotheses that are organized into a decision tree and the system then searches through its vocabulary for words that match the hypotheses. In the next sections, we will consider whether humans employ this kind of higher-level reasoning to recognize basic sounds.

2.4 Human speech recognition

The brain tends to automatically categorize speech sounds. Experiments conducted by House [1962] showed that when listeners hear nonspeech synthetic sounds which are gradually made more speechlike, an abrupt boundary is found, where on one side the sounds are perceived as speech and on the other side they are not. Further, in experiments by Liberman [1957] a phoneme in synthesized words was gradually varied until it became another phoneme, for example "bad" to "gad" and "pit" to "bit." Listeners did not hear this gradual variation, but instead made sharp

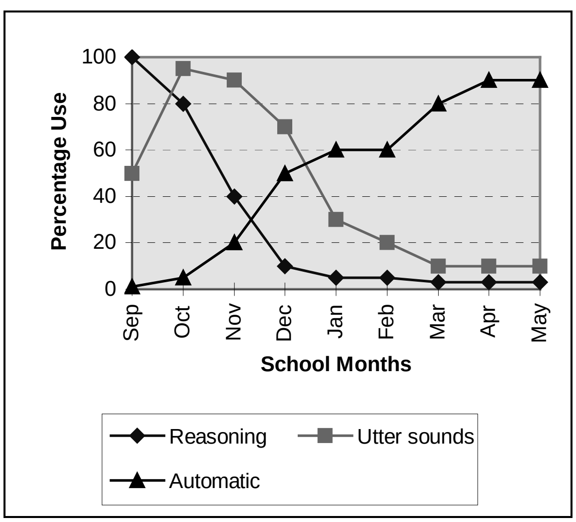

Table 2. Averaged results from retrospective survey. Subjects were asked to plot, for each month, the percentage of time they employed each modality in learning the sounds of a second language.

distinctions among initial consonants. It appears that there is an innate predisposition in the human nervous system to immediately classify speech inputs. Certainly initial training is required to develop these explicit categories, but once present, sound recognition seems an automatic process.

2.5 Students learn phonemes

We recently interviewed and surveyed students learning a foreign language that had a significantly different sound system from their native language. In particular, these were American English speakers learning the Korean language over two consecutive school semesters. The students began the school year with only knowledge of English.

Two students were interviewed separately. The students: (1) described a time during which they had to train their ear to recognize the sounds of the new language, claiming that at first they engaged a lot of reasoning and predictive knowledge to identify sounds (which we refer to as "reasoning"); (2) claimed they uttered competing words in their minds to compare different sounds (which we refer to as "uttering"); and (3) agreed that once the sounds became familiar to their ear, detection appeared to become automatic even if the meanings of the words were unknown (which we refer to as "automatic").

We then surveyed four students and asked them to retrospectively plot how much they used: reasoning, uttering, and automatic detection over the school year. The students plotted values for each month, for each of the three curves. The averaged results are shown in Table 2 and can be interpreted as follows. When the first term began, reasoning was relied on since students could not distinguish Korean sounds at all; context and higher-level knowledge were critical. Several Korean phonemes sound similar, such as /m/ and /n/ in English, so the students often uttered competing sounds internally. While speech perception relies on the listener's familiarity with the language, the talker's characteristics and more, the recognition of speech sounds appeared automatic once the sound categories were established.

In the previous survey we can draw analogies to all three approaches in speech recognition. Acoustic-phonetic recognition, where one reasons about the sound, is similar to the deliberation the students undertook when contemplating what sounds they heard. Likewise, the stochastic approach, as was spawned by Harpy, where one makes comparisons to known sounds and has expectations of what sound is next based on known words, is similar to the uttering practice described by the students who compared similar sounds in their heads. Our goal with the BeBe System is to model the human ability to directly categorize sound in real time without higher-order reasoning. The students described employing this behavior once they had acquired the ability to categorize the speech sounds of the new language. The recognition of the sounds became automatic.

Even though these survey results may not be statistically or psychologically persuasive, they do agree with the findings of Fletcher and his colleagues, who studied the principles behind human speech recognition from 1918 to 1950 at Bell Labs [Allen, 1994] and concluded that humans decode speech sounds into independent units at an early stage, before semantic context is used.

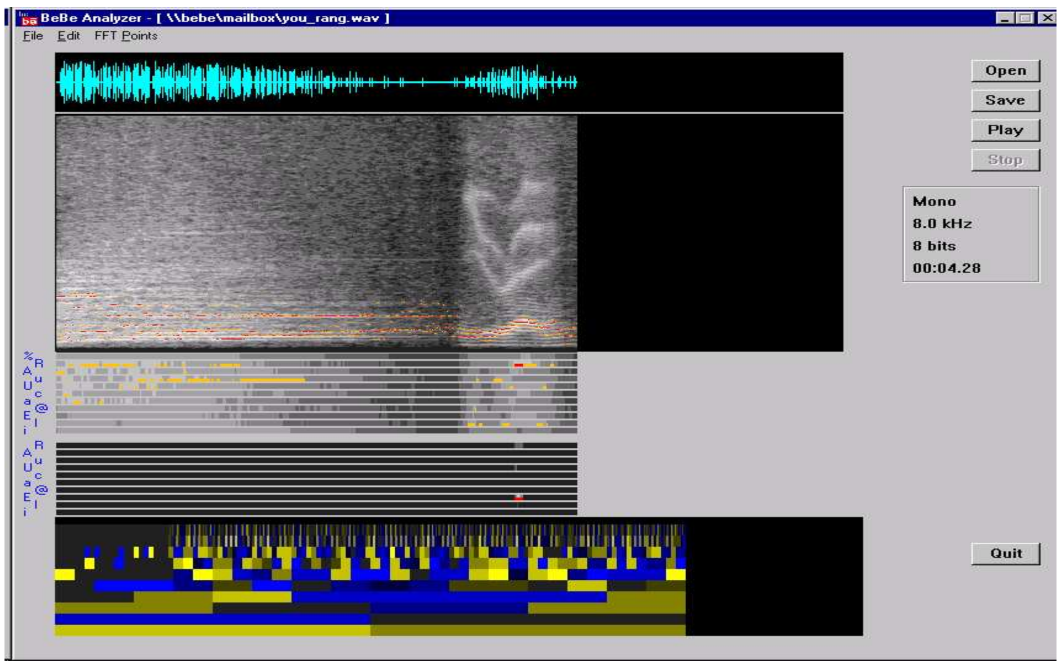

Diagram 2. A presentation of BeBe$_0$ at work. There are five display areas, from top to bottom: the waveform, the FFT spectrogram, the results from the detection algorithms, other information from the detection algorithms, and a wavelet transform.

3 Design

We sought to model the human ability to automatically categorize speech sounds since this did not require a semantic model of the language. Instead it involved an almost "biological" detection of basic sounds directly from the waveform. Further, we wished to evaluate the Scrub architecture when applied to the well-suited problem of phoneme detection.

For each phoneme, or basic sound category, there are one or more detection algorithms, each with precedence based on the number of samples that constitute the algorithm's assigned sound; detectors for phonemes with longer duration have higher precedence. For each sample in the sound the detection algorithm with the highest precedence reporting the greatest likelihood above a threshold value is considered to have identified an instance of its phoneme. In development, detection algorithms can be executed sequentially in order of precedence to avoid parallel execution. Diagram 1, presented earlier, provides an overview. All detectors report their findings to the polling algorithm, which then makes final decisions based on the certainty of the detectors, their precedence, and time analysis of past results.

Often a separate detection algorithm is used for word-final than for word-initial occurrences of consonants in words since systematic relationships exist between phonemes in particular positions within a word and their pronunciation. Consider the [p] and [p^h] allophones of /p/ presented earlier. Some detectors classify information about background noise. These detectors continuously report findings. There are also special detectors like those that determine whether human speech is present and having these results reduces the number of false positives. At run-time the user can set the threshold and use of these special detectors.

Knowing what instances have already been found in the sound can be useful in reducing ambiguity. For example, if the system encounters the phoneme sequence < /i/ /I/ /i/ > in contiguous sound samples, where the duration of each phoneme is less than its typical length, then the /i/ result can prevail with more certainty and account for all these samples. This type of drift in the speaker's voice is common. Similar sequence inconsistencies occur due to coarticulation at word and syllable boundaries. Thus, maintaining a queue of recent detector results allows the polling algorithm to summarize a series of findings as being a single instance of one phoneme.

The queue used by the polling algorithm can grow to hold a one-second history or about 100 detector results. In producing a phonetic transcript, the polling algorithm outputs one instance of a phoneme for roughly 20 or more contiguous results. Detectors primarily base their decisions on frequency information while the polling algorithm resolves further uncertainty using time analysis.

4 Implementation

We implemented a version of the BeBe System, termed $BeBe_0$, in C++ on a Pentium Pro processor machine running at 200 MHz with 32MB RAM and executing the Windows 95 operating system. A Shure SM10A microphone was used in a SoundBlaster-equivalent audio card with software adapted from Rimmer [1995]. The recording environment was an office setting in which background noise included a loud ventilation system and the electronic hum of three computer systems. The sound pressure level was estimated at 63 dBA where a typical office with a single computer is around 45-50 dBA [Rabiner, et al., 1993].

Here is a walk through the system. A forward-sliding window identifies 512-sample segments of the audio waveform, nominally sampled at 22050 Hz, monaurally, with 8-bit resolution. Spectral slices are developed by passing these segments through a Hamming window to an FFT analyzer. The moduli of the transformed values are approximately weighted according to the frequency response of the human ear [Jones, 1993] and smoothed along the time axis by a simple recursive filter.

With these steps completed, each detection algorithm has available to it a vector of the sample segment, and vectors of the FFT before and after weighting. The average spectral density for the weighted FFT is also computed (by averaging the values in the vector up to the Nyquist frequency) and posted.

Each detector is free to use any of this information to determine its certainty that the sample is an instance of its phoneme. The certainty factor can be conceived as a number between 0 and 1. For many of the vowels, we found that simple heuristics concerning typical duration and templated spectral-band densities for formants F1, F2, and F3 [Peterson and Barney, 1952] could accurately identify instances. False positives were gated out by requiring a minimum average spectral density (a loudness threshold) and then a minimum ratio of template-band density to average density for suspected vowels (a minimum vowel-signal-to-noise). Algorithm 1 provides a

Given a weighted, 512-segment FFT *wfft*, its average spectral density d_{avg}, and sampling frequency 22050 Hz, the response of the /R/ detector is determined as follows:

a. The computation of the band spectral density d_{band} is based on the averaged formants for /R/ from Peterson and Barney [1952]. The index *i* of a frequency *F* in vector *wfft* is determined by:

$$i = \tfrac{512}{22050} F \, .$$

Then,

$$d_{\text{band}} = \underbrace{\sum_{i=10}^{11} wfft[i]}_{\text{F1}\in[490\pm 23]} + \underbrace{\sum_{i=30}^{32} wfft[i]}_{\text{F2}\in[1350\pm 29]} + \underbrace{\sum_{i=38}^{39} wfft[i]}_{\text{F3}\in[1690\pm 29]} \, .$$

b. If $d_{\text{band}} < 50$, then exit and return 0 since the minimal speech signal is not present.

c. Let $b = (d_{\text{band}} - 3d_{\text{avg}})$. If $b \leq 0$ then return 0; else, let $b = \tfrac{1}{6} b d_{\text{avg}}$. If $b > 1$ then let b=1. Return b as the certainty factor.

Algorithm 1. The non-adaptive version of the /R/ detector as used in BeBe_0. An example of the /R/ sound is found in st<u>ir</u>. The magic values 50, 3 and 6 were determined empirically.

listing of the /R/ detector that uses this simple strategy. Other detectors work differently, and some could be made adaptive.

BeBe_0 detects formants based on fixed templates. A logarithmically varying template based on critical bands [Schroeder, et al., 1979], such as the Bark scale used by Huang [1991], would likely improve detector accuracy and speaker independence in BeBe.

BeBe_0 gets robust results with only the fixed-template version of the /R/ detector. If one were to build a BeBe System with only /R/ and /U/ detectors, where the /U/ detector is similarly constructed, the number of correct detections for /R/ when present would remain high, but the number of false detections, where /R/ is detected but not present, would also be very high. Consider the definitions for correctness and accuracy listed in Definition 1. In the BeBe architecture, having detectors that are somewhat exclusive and correct, even if they are not accurate, appears to boost the performance of the overall system.

Diagram 2 shows BeBe_0's detectors at work. From top to bottom are displays of the voltage waveform, the FFT-based spectrogram, two ribbon-sets of results, and a wavelet transform. The sound in the waveform is a loud cymbal crash and then the words “you rang.” Of particular note is the upper ribbon-set display, showing the vowel detection for each Arpabet symbol on the left-hand side. The monochrome intensity in each ribbon shows that detector’s certainty that it has identified its vowel. During the cymbal crash, the false detections are overruled, since the average spectral density exceeds the band densities of any of the vowels. On a color display, such determinations show as orange ribbons. Later in the waveform, during the words “you rang,” the selectivity of the detectors is more apparent; for example, the /R/ line is bright, denoting the presence of the /R/ phoneme. A red ribbon indicates which detector has the highest certainty at points along the waveform. The second ribbon-set display shows a competing group of detectors, with their heuristics still under development, and some of these detectors use information from the wavelet transform.

$$\text{correctness} = \frac{\text{number correct}}{\text{total occurences}}$$

$$\text{accuracy} = \frac{\text{number correct} - \text{false positives}}{\text{total occurences}}$$

Definition 1. The BeBe architecture favors independent detectors that are correct but not necessarily accurate.

5 Results

We conducted an experiment to determine how well BeBe$_0$ would identify phonemes in discrete and continuous speech. The subjects were 4 male adults, all native American English speakers, each from a different state: Alabama, Illinois, California and Washington. Each subject was given a page containing 5 lines of text. The first 2 lines had isolated words and the subjects were instructed to pause between each word. The last 3 lines were full sentences and the subjects were to read them aloud in a natural manner. Table 3 contains a copy of these words. One subject did not speak 1 word and another did not speak 2 of the words in the text, so these were not counted.

Bob. Hat. Heard. Think. Thought. Fee.
Head. Book. Boot. Mud. Through.
Kim needs that book first.
She saw the fat blue bird.
His red boot got muddy.

Table 3. The lines above were spoken into BeBe$_0$ by the subjects. The first 2 lines have a pause between each word and the last 3 are continuous sentences.

BeBe$_0$ was tested on its recognition of 4 phonemes: /@/ as in bat, /R/ as in stir, /I/ as in fit, and /U/ as in should. These phonemes were chosen because they cover each vocal-tract articulatory configuration for vowels: /I/ and /@/ are considered front vowels, /R/ is a middle vowel and /U/ is a back vowel.

	R	A	u	U	c	a	@	E	I	i
/R/	**12**									
/A/		5				1				
/u/	1		9	1						
/U/		**1**		**7**						
/c/				1	8				1	
/a/						6	1			
/@/					**1**	**1**	**11**			
/E/	1							7	1	
/I/	**1**								**10**	**1**
/i/			1						1	10

Table 4. One review of results from all 10 detectors in BeBe$_0$. Each row denotes a phoneme that occurred in the sound, and each column displays the results from a detector. If the results were perfect, all numbers would appear in the diagonal. The shaded rows highlight the phonemes under study, and is also where both reviews agree.

A total of 10 competing phoneme detectors were used, and these were based on the averaged formants from Peterson and Barney. Tables 4 through 8 have summary results. No training was performed. During the development of BeBe$_0$, we recorded approximately 50 three-second recordings of various sounds and speakers. These files were useful during debugging, but none of these recordings included the test lines shown in Table 3. Samples for one subject were included, though the subject did not speak the text in Table 3.

The results from all 10 detectors were scored in 2 different reviews, which both agreed on the correctness of the 4 phoneme detectors shown in Table 7, but differed with respect to the correctness of the other detectors. The source of the problem was in transcribing the sounds from the speakers. For example, some subjects pronounced *the* with an /i/ sound while others used an /E/ sound. Table 4 shows a summary of one review. The other review differs in about 7 instances.

Considering the 4 phonemes under study, $BeBe_0$ performed quite well, from 83-100% correct. There was no appreciable difference between isolated and continuous speech. There were a total of 25 phonemes in the text and 4 speakers, totaling 100 possible detections. The detectors reported false positives as follows: /@/ had 1, /R/ and /I/ each had 3 and /U/ had 2, which is 1-3%.

Phoneme	Total Occurrences	Correctness
/R/	**12**	**1.00**
/A/	6	0.83
/u/	11	0.82
/U/	**8**	**0.88**
/c/	10	0.80
/a/	7	0.86
/@/	**13**	**0.85**
/E/	9	0.78
/I/	**12**	**0.83**
/i/	12	0.83

Table 5. Correctness results for data in Table 4. The average correctness overall is 85%.

Detector	Totals	False Positives	Accuracy
R	**15**	**3**	**0.75**
A	6	1	0.67
u	10	1	0.73
U	**9**	**2**	**0.63**
c	9	1	0.70
a	8	2	0.57
@	**12**	**1**	**0.77**
E	7	0	0.78
I	**13**	**3**	**0.58**
I	11	1	0.75

Table 6. Accuracy results for data in Table 4. The average accuracy overall is 69.5%.

	#correct	Total	%correct
/@/	11	13	85
/R/	12	12	100
/I/	10	12	83
/U/	7	8	88

Table 7. Results from human subjects speaking the text in Table 3 with $BeBe_0$ detecting these phonemes. For a given phoneme, %correct = (number identified correctly) / (total occurrences) $\times$ 100.

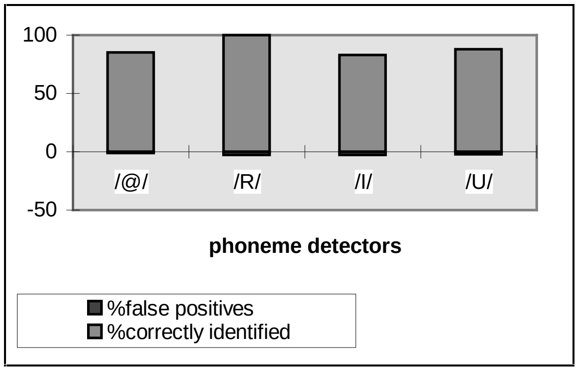

Table 8. Data from Table 7 with false positives.

5.1 Comparison of others

As demonstrated in the previous subsection, evaluating the performance of a speech-based system is quite difficult and comparing different systems is even worse. Different systems are meant to accomplish different goals, and by design perform better on some tests than others. Also, we cannot compare the results of the simple experiment discussed in the previous subsection, whose purpose was to illustrate the plausibility of the BeBe approach, with experiments that report results based on a large corpus over many speakers. Nevertheless, looking at the reported accuracy of other systems helps evaluate the motivation for the BeBe System.

Most of the early experiments focused on vowel recognition. The summary of results listed in Table 9 implies that recognizing vowels is at least as difficult as identifying consonants. Schmid [1996] reports 62% accuracy using a unigram language model and detecting phonemes /t/ and /m/ using a vowel classifier based on the N-best consistent interpretations of formant information.

The results in Table 9 for vowel accuracy are based on the TIMIT corpus, which contains speech from 630 speakers from 8 major dialects of American English, each speaking 10 sentences. As future work, $BeBe_0$ and its adaptive version should be tested on TIMIT sound samples to allow direct comparison.

Type	Researcher	Description	Accuracy
vowel	Meng, et al. [1991]	auditory model	64.5%
vowel	Carlson, et al. [1992]	neural net, gender info	65.6%
phoneme	Chigier, et al. [1992]	neural net	78.0%
phoneme	Digalakis [1993]	male only	73.9%
unigram	Schmid [1996]	n-formants, detect /t/ and /m/ phonemes	61.1%
unigram	Schmid [1996]	n-formants with cepstral	62.0%

Table 9. Accuracy results from various researchers. Vowel findings are based on the TIMIT corpus. The unigram language involves testing with isolated phoneme sounds.

Perceptual experiments using the TIMIT database were conducted by Cole and Muthusamy [1992]. Vowel identification for the 16 TIMIT vowels when isolated from sentences and then played to human subjects had a correctness of 0.55, which rose to 0.66 when acoustic context was provided. Machine recognition of vowels already exceeds this correctness level, which suggests that training of the human auditory system is sensitive to statistical correlation of related local context beyond what was tested. Some consonant-vowel and vowel-consonant transitions, for example, are common and others do not occur at all and such may be related to how humans pronounce and recognize nonsense words [Sweeney, 1996b].

6 Discussion

We have presented the BeBe architecture and demonstrated its robustness. There are about 40 phonemes in English and we expect about 100 detection algorithms in BeBe; though many of them are not yet active, the feasibility of this approach has been illustrated even though our experiment did not utilize the BeBe architecture’s ability to employ different algorithms for different phonemes. In closing, we will discuss how BeBe might be used in a speech system, examine other phoneme recognizers, and assess BeBe’s performance and potential.

6.1 BeBe as the first stage

Earlier we discussed the problem of word boundaries where the identification of a word even from a proper phonetic transcription requires knowledge of local context at a higher level. We could easily imagine a greedy algorithm that starts processing phonetic symbols matching the best guess at the word so far; if that does not lead to identification of the word and the second word also, then backtrack and take the next best guess at the word. Or, we could employ a statistical or HMM approach to the results from BeBe and the resulting system would be far simpler than traditional HMM models.

Another strategy applies the BeBe architecture at a higher level. A phonetic transcript is sent to a second stage that converts the transcript to poorly spelled English words using sound-to-spelling rules, phoneme-to-grapheme mappings [Hall, 1961], or orthographic rules [Sweeney, 1996]; see Diagram 3. The candidate words from the second stage then go to a spelling corrector which produces the final result. The entries at all stages include likelihood measures, and the most likely entries (no more than seven) are passed to the next stage. The most attractive aspects to these approaches, not including the spelling corrector, are: no large vocabulary is needed for storage and retrieval, and the number of words recognizable is infinite and includes nonsense words such as “throck” and “zat” as well as proper names and terms specific to a profession or field with no additional training. The recognizer, without the spelling corrector, can spell any word that is pronounceably consistent to the language’s pronunciation rules, and with the spelling corrector this can be limited to only the list of known spellings. Also, these approaches adhere to well-defined abstraction layers that have no global methods, and all three approaches to the middle stage could be applied in parallel.

As part of their study of human speech recognition, Fletcher and his colleagues proposed a model of human speech recognition that consisted of a cascade of recognition layers, starting with the cochlea [Allen, 1994]. As in the BeBe architecture, no feedback is assumed between layers; Fletcher’s abstractions are similar. The first layer, based on the cochlea, determines the signal-to-noise ratio in about 2800 overlapping critical band channels; the second layer extracts about 20 speech features from the channel information in a local manner; the next layer maps those features onto phonemes, and then final layers determine syllables and words. In the next sections, we discuss other phoneme recognizers and then assess BeBe’s performance.

6.2 Recognizing vowels

Ironically, speech recognition research began with vowel and phoneme detection. Vowels are typically long in duration compared to consonants and are spectrally well-defined so the presence of one can be reliably detected [Rabiner, 1993], but identifying the presence of a specific vowel varies from speaker-to-speaker and is more difficult as was shown in Table 9. In 1959, researchers at MIT Lincoln Laboratories [Forge, et al.] built a detector capable of recognizing a few vowels contained within a /b/-vowel-/t/ pattern. The system used a filter band analyzer and an estimate of time duration. In 1961, researchers in Japan [Suzuki, et al.] built a hardware vowel recognizer for some Japanese vowels which used an elaborate filter bank spectrum analyzer along with logic that connected the weighted outputs of each channel to a vowel-decision circuit where a majority-decision logic scheme was used to identify the spoken vowel. In 1962 another group of researchers in Japan built a hardware phoneme recognizer [Sakai, et al.] which provided gross categorization using a segmenter along with zero-crossing analysis. In 1966, Reddy pioneered research to dynamically track phonemes for continuous speech recognition.

Today's speech recognition systems still rely heavily on vowel recognition to achieve high performance, but the shift towards statistical processing spawned by Harpy in the 1970s pushed researchers to incorporate higher-level knowledge and accept gross uncertainty at the acoustic-phonetic level or ignore it altogether. In concluding, we return our attention to the BeBe System in light of these works.

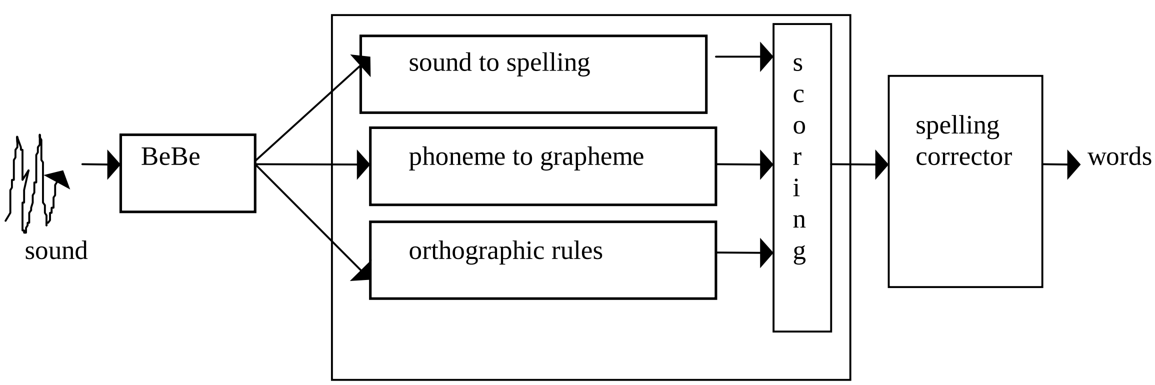

Diagram 3. Block diagram for a speech recognizer that uses the BeBe System as the first stage. The middle stage has a design similar to that of the BeBe System, with three parallel processes, where each process computes spellings from the phoneme-like sequence that's outputted from BeBe. Associated with each of these spellings is a likelihood measure as to how certain the process is that the phoneme sequence represents that spelling. The top few most likely spellings are then sent to a spelling corrector where the most likely words are determined and outputted.

6.3 Assessment and future work

Consider the phoneme detectors for /@/, /R/, /I/, and /U/ discussed earlier in the Implementation and Results sections. $BeBe_0$ used FFT spectrum analysis and weights, time analysis, and a voting scheme to identify phonemes. This is similar to the technology employed by researchers in the 1950s and 1960s for building recognizers, though admittedly $BeBe_0$'s implementation is more powerful and takes advantage of improved computational prowess. Some of its power comes from its simplicity.

Having each detector report a certainty factor eliminates complicated and conflicting hierarchical rules which would otherwise be inevitable as the system would expand to include more detectors. Unlike previous recognizers and current classifiers, where the same procedures are applied to the recognition of all phonemes, detectors in the BeBe architecture are independent and can exploit other methodologies. For example, pitch, spectral tilt, relative formant amplitude and wavelet analysis were not necessary for detecting phonemes in $BeBe_0$ but may prove quite useful in the identification of other phonemes. The BeBe architecture also supports the use of special detectors to identify ambient noises, and perhaps eventually, to isolate different speakers. Assigning one or more detectors to each phoneme allows the recognition algorithms to remain simple and the overall result produces well-modulated abstractions. Complexity in $BeBe_0$ was governed by the required number of multiplications for the FFT which was $O(n \lg n)$, where $n = 2^k$ is the window size. This is constant during operation and spatial complexity is likewise constant.

The phoneme detectors in $BeBe_0$ identified vowels but many of the same spectral tools will be useful in identifying consonants as well [Weinstein, et al., 1975]. Further, the results in Table 9 suggest that machines can detect consonants more accurately than vowels.

Unfortunately, each of the detectors in $BeBe_0$ were hand-coded and this conflicts with the adaptive nature of human auditory perception. We believe however that coding detectors by hand will provide a set of tools from which a representation will emerge that can be used to computationally "evolve" detectors. In such a case, the resulting BeBe System would hear sounds and then build its own detectors, thereby completely emulating the auditory learning behavior described by the Korean language students.

References

[Allen, 1994] Allen, J. How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*: 2(4).

[Baker, 1975] Baker, J. Stochastic modeling for automatic speech understanding. *Speech Recognition*, edited by Raj Reddy. London: Academic Press: 521-541.

[Bourlard, et al., 1996] Bourlard, H., Hermansky, H., and Morgan, N. Towards increasing speech recognition error rates. *Speech Communication*: 18(3).

[Carlson, et al., 1992] Carlson, R. and Glass, J. Vowel classification based on analysis-by-synthesis. *Proceedings of ICSLP*, 575-578.

[Chigier, et al., 1992] Chigier, B. and Leung, H. The effects of signal representations, phonetic classification techniques. *Proceedings of ICSLP,* 97-100.

[Cole, et al., 1980] Cole, R.A., Rudnicky, A.I., Zue, V.W. and Reddy, D.R. Speech as patterns on paper. *Perception and Production of Fluent Speech*, edited by R.A. Cole. Hillsdale: Lawrence Erlbaum Associates: 3-50.

[Cole, et al., 1992] Cole, R. and Muthusamy, Y. Perceptual studies on vowels excised from continuous speech. *Proceedings of ICSLP,* 1091-1094.

[Digalakis, et al., 1993] Digalakis, V., Rohlicek, J., and Ostendorf, M. ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition. *IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* 431-442.

[Erman, et al., 1980] Erman, L., Hayes-Roth, F., Lesser, V. and Reddy, D.R. The Hearsay-II speech-understanding system: integrating knowledge to resolve uncertainty. *Computing Surveys*, 12(2), 213-251.

[Forgie, et al., 1959] Forgie, J. And Forgie, C. Results obtained from a vowel recognition computer program. *Journal of Acoustic Society of America*, 31(11):1480-1489.

[Hall, 1961] Hall, R.A. *Sound and spelling in English.* Philadelphia.

[House, 1962] House, A.S. On the learning of speechlike vocabularies. *Journal of Verbal and Learning Verbal Behavior*, 1, 133-143.

[Huang, 1991] Huang, C.B. *An acoustic and perceptual study of vowel formant trajectories in American English.* Massachusetts Institute of Technology, Research Laboratory of Electronics: Technical Report 563.

[Jones, 1993] Jones, E.R. *Contemporary College Physics*, 2^{nd} ed., Reading: Addison-Wesley: 426.

[Klatt, 1977] Klatt, D. Review of the ARPA Speech Understanding Project. *Journal of the Acoustic Society of America*, 62, 1345-1366.

[Kornfeld, 1979] Kornfeld, W.A. *Using parallel processing for problem solving*. Massachusetts Institute of Technology, Artificial Intelligence Laboratory: Memo 561.

[Liberman, 1957] Liberman, A.M. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54, 358-368.

[Lippmann, 1996] Lippmann, R. Recognition by humans and machines: miles to go before we sleep. *Speech Communication*: 18(3), 247-248.

[Markowitz, 1996] Markowitz, J. *Using speech recognition*. Upper Saddle River: Prentice-Hall: 129-135

[Meng, et al., 1991] Meng, H. and Zue, V. Signal representation comparison for phonetic classification. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 285-288.

[Peterson and Barney, 1952] Peterson, G.E. and Barney, H.L. Control methods used in a study of the vowels. Journal of the Acoustical Society of America, 24(2), 175-184.

[Pieraccini, et al., 1991] Pieraccini, R., Lee, C., Giachin, E. and Rabiner, L. Complexity reduction in a large vocabulary speech recognizer. *IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 729-732.

[Rabiner, et al., 1993] Rabiner, L. and Juang, B. *Fundamentals of speech recognition*. Englewood Cliffs: Prentice-Hall: 392.

[Reddy, 1966] Reddy, D. *An approach to computer speech recognition by direct analysis of the speech wave*. Stanford University, Computer Science Department, Technical Report C549.

[Rimmer, 1995] Rimmer, S. *Advanced multimedia programming.* New York: Windcrest/McGraw-Hill.

[Sakai, et al., 1962] Sakai, T. and Doshita, S. The phonetic typewriter, information processing. *Proceedings IFIP Congress.* Munich. Also discussed in [Rabiner, et al., 1993].

[Schmid, 1996] Schmid, P. *Explicit, N-best formant features for segment-based speech recognition*. Oregon Graduate Institute, Department. of Computer Science, Technical Report CSE-96-TH-003.

[Schroeder, et al., 1979] Schroeder, M.R., Atal, B.S., and Hall, J.L. Objective measure of certain speech signal degradations based on masking properties of human auditory perception.

Frontiers of Speech Communication Research, edited by B. Lindblom and S. Ohman. London: Academic Press: 217-229.

[Suzuki, et al., 1961] Suzuki, J. and Nakata, K. Recognition of Japanese vowels – preliminary to the recognition of speech. *Journal Radio Research Laboratory* : 37(8):193-212.

[Sweeney, 1996a] Sweeney, L. Replacing Personally-Identifying Information in Medical Records, the Scrub System. In: Cimino, JJ, ed. Proceedings, *Journal of the American Medical Informatics Association.* Washington, DC: Hanley & Belfus, Inc, 1996:333-337.

[Sweeney, 1996b] Sweeney, L. Automatic acquisition of orthographic rules for recognizing and generating spellings. MIT. AI Working Paper.

[Weinstein, et al. 1975] Weinstein, C., McCandless, S., Modshein, L., and Zue, V. A system for acoustic-phonetic analysis of continuous speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing,* 23(1), 314-327.

[Zue, 1985] Zue, V. W. The use of speech knowledge in automatic speech recognition. *Proceedings of the IEEE,* 73(11), 1602-1615.