STATISTICAL SPEECH RECONSTRUCTION AT THE PHONEME LEVEL

Michael Savic, Michael D. Moore, Christopher Scoville

ECSE Department
Rensselear Polytechnic Institute
Troy, New York 12180, USA

ABSTRACT

Statistical methods for reconstructing speech at the phoneme level are used to find missing phonemes that are removed from sentences in the TIMIT corpus. Probabilities for the occurrence of the missing phoneme(s) are generated and the most likely candidate(s) selected to reconstruct the sentence. Method includes symmetric and asymmetric 'confidence windowing' around the missing phoneme(s) for determination of the most likely candidates. Reconstruction rates for one or more phonemes missing in a sequence can exceed 85%.

INTRODUCTION

Recently the Signal Research Group at RPI finished a major project that involved automatic language identification. The recognition was based on *language parameters* such as transition probabilities from phoneme to phoneme, observation probabilities, durations of phonemes, and others. The idea for this research project involves the reverse procedure - missing parts of speech can be reconstructed if the language is known. In other words, the missing phonemes can be reconstructed from *language parameters* of a given language. These parameters include the observation and transition probabilities of phonemes, the duration of phonemes and others. It is often the case in speech communications that the received speech is damaged due to some kind of interference. The objective of our research is to use statistical methods to reconstruct the damaged speech by replacing the missing phoneme(s). We consider a sequence of phonemes (symbols) as illustrated in Figure #1. The picture shows a word, i.e., a stream of phonemes with removed symbols, the identities of the removed symbols are unknown. The objective of our research is to use statistical methods to reconstruct the damaged stream by replacing the missing phoneme(s) (symbols). Preliminary statistical analysis of phonetic streams (TIMIT sentences) is done with the use of several confidence window templates that are used to bracket the location of the missing phoneme(s), and provide statistics on the most likely candidates for the missing phoneme(s). Initially used confidence windows are illustrated in Figure #2.

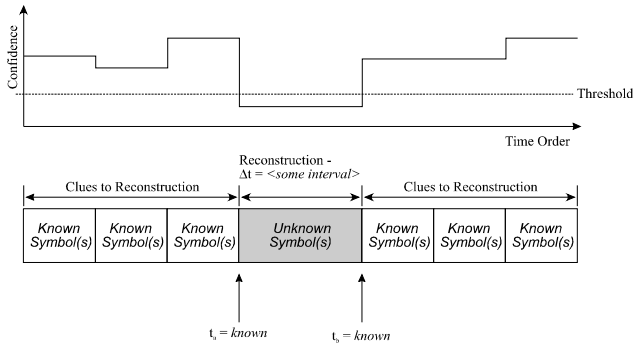


Figure #1. The general problem of reconstruction on a phoneme basis. Given a stream of phonemes from the TIMIT corpus, reclaim the missing phonemes using statistical techniques.

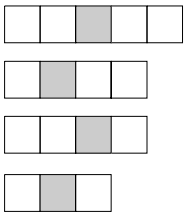


Figure #2. Initially used confidence window templates to ascertain most likely candidates. Missing phoneme is contained in gray box; phonemes in white boxes are used to determine probability for missing phoneme.

The TIMIT database consists of 6300 sentences (some of which are repeated), comprised of the standard 61 phonemes.

1. SINGLE PHONEME RESULTS

A procedure was employed for single, missing phonemes in which a confidence window of preset size was used to accumulate observation and transition probabilities for a particular phoneme in the sequence that surrounds the missing phoneme. From the knowledge of probabilities of occurrences of particular phonemes in this confidence window, most likely candidates for the missing phoneme were retrieved. The sentences of Figure #4 were chosen at random from the TIMIT database, as were the removed phonemes.

Results from this procedure for the single missing phoneme case are shown in Figure #4. Here we illustrate the overall results for the three-element window (Figure #3), a known phoneme on either side of the unknown one. The raw recognition rate for this confidence window was 33.91% (Figure #5).

Figure #3. Three-element confidence window used for determining the candidate phonemes of Figure #4.

	TIMIT Sentence	Missing Phoneme	Candidates in Order	P(R), Raw
1	SI2048.PHN	z	ax, z, ix	0.25
2	SI720.PHN	t	t, sh, ch	0.72
3	SI1088.PHN	ix	ix, aa, ux	0.169014
4	SI2129.PHN	f	z, v, f	0.333333
5	SI509.PHN	ey	ey, iy, ix	0.777778
6	SI1477.PHN	eh	ih, ix, eh	0.106383
7	SX325.PHN	n	n, l, th	0.25
8	SX145.PHN	ng	ng, n, z	0.357143
9	SI1789.PHN	b	b, s, el	0.949153
10	SI617.PHN	n	n, z, tcl	0.357143
11	SI598.PHN	ix	ix, uh, ih	0.314815
12	SI1968.PHN	q	q, l, hh	0.652174
13	SI1409.PHN	n	w, m, n	0.115385
14	SI1831.PHN	dcl	dcl, tcl, q	0.959064
15	SX87.PHN	iy	iy, ay, aw	0.172414
16	SI1585.PHN	ch	t, ch, d	0.109489
17	SI2128.PHN	epi	eh, epi, aa	0.181818
18	SI1111.PHN	ux	ax, ux, pau	0.166667
19	SI1147.PHN	s	s, w, dh	0.441176
20	SI1178.PHN	z	w, sh, bcl	0
21	SI485.PHN	jh	jh, z, ix	0.727273
22	SI1581.PHN	axr	eh, ae, ux	0
23	SI1289.PHN	m	m, tcl, z	0.333333
24	SI1848.PHN	ay	ih, ix, ey	0
25	SI730.PHN	pau	bcl, pau	0.034483

Figure #4. Results from the application of the window shown in Figure #3 to TIMIT sentences.

The raw accuracy of the three-element window above was calculated by taking the total number correct phonemes divided by the total number of candidate phonemes returned. In Figure #5 we calculate an additional measure, the ratio of correct phonemes to the population of the top 3 candidates returned.

The remaining windows of Figure #2 were applied to the chosen sentences and the results are illustrated in Figure #5. Of note is the increasing reconstruction rate with additional known phonemes surrounding the unknown phoneme, but at the expense of referencing the very same sentence in TIMIT from which the missing phoneme was taken. Additionally, the 'screening' effect of the larger confidence windows is illustrated by comparing the raw accuracy with the rate calculated by dividing the number of correct phonemes by the population count of the top 3 candidates. Larger groups of phonemes before and after the missing phoneme will usually not represent transitions from the last known phoneme to the most probable unknown phoneme, unless they are within the same (or very similar) word. Thus, increasing the known information in the confidence window tends to lead toward larger language units (i.e., words). If the window spans words, reconstruction tends towards a specific sentence. These results are illustrated in Figure #5. Here it is seen that the larger confidence windows discriminate possibilities to a greater extent over the smaller confidence windows.

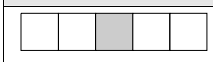
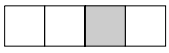
Confidence Window	Reconstruction Rate, Raw	Reconstruction Rate, Top 3
	85.25%	85.25%
	58.52%	62.33%
	66.16%	68.31%
	33.91%	44.55%

Figure #5. Additional confidence window templates and the reconstruction rates, raw and top 3 candidates.

2. MULTIPLE PHONEME RESULTS

Runs with the program of Figure #6 were performed to investigate the use of multiple confidence window types on multiple missing phonemes. With this procedure, a confidence window is chosen, as well as a phoneme to remove to damage the speech stream. Reconstruction is performed, and another window is chosen along with

another symbol to damage in the same sentence. This procedure is then repeated the desired number of times (here five damage incidents) :

Figure #6. A picture of the application interface used for investigating multiple missing phonemes.

The sentence chosen for multiple phoneme reconstruction is illustrated in Figure #7, sentence *\timit\test\DR7\MDLF0\SX53.PHN* ('Even a simple vocabulary contains symbols') from the TIMIT corpus, before and after being damaged (the damaged symbols are question marks) :

```
Q iy v ix nx IX S ih m pcl p el V
ow kcl k ae ux l ax r iy KCL k ix n
tcl t ey n s pau S ih m bcl b el s

Q iy v ix nx IX S ih m ? p el V ow
kcl k ae ux l ax r ? KCL k ? n tcl
  ? ey n s pau S ih ? bcl b el s
```

Figure #7. Sentence to reconstruct, before (top) and after (bottom) damage.

Figure #8 illustrates the windows used and the reconstruction results P(R) for the regions of damaged speech. Here is seen that a combination of the two illustrated windows is 100% correct in reconstructing the damaged sentence of Figure #7. The specific window used in each reconstruction can be inferred from the size of the phoneme string containing the (?) character. A phoneme located at the start of a word is illustrated in capital letters :

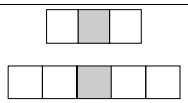
	?	✓	P(R)
ih m ? p el	pcl	pcl	1.0
r ? KCL	iy	iy	1.0
k ? n	ix	ix	1.0
n tcl ? ey n	t	t	1.0
S ih ? bcl b	m	m	1.0

Figure #8. The results of multiple phoneme reconstruction; the missing pattern in the confidence window (?), the correct missing symbol (✓), the chosen symbol, and the reconstruction probability (columns 1 to 4) for the damaged sentence of Figure #7.

The speech parameters necessary for word reconstruction can be obtained analytically using the Hidden Semi Markov Model [2]. These parameters are language dependent. The HMM suffers from the limitation that it inaccurately models the state durations of phonemes (occupancy distributions) as geometric. The difficulty appears when each state represents a phone or group of phones; if transitions to the same state are possible, the time spent in each state is a random variable with a geometric probability mass distribution as :

$$\Pr(\text{time spent in state} = n) = p^{n-1}(1-p)$$

where p is the probability of remaining in the same state. Analysis in [2] of the TIMIT database clearly shows a geometric representation as inaccurate.

The HSMM is similar to the HMM, except that in cases the HSMM does not necessarily enter a new state each clock cycle. This is because of differing (and finite) durations of the phonemes represented by each state (a clock cycle is 'shorter' than the residence time in any state or phoneme duration, being a speech sample Δt or group of samples in a frame, etc). Formally, a HSMM is a more general class of Markov chains where the state occupancy is defined by arbitrary probability mass distributions, in these semi-Markov chains the Markov property is not necessarily satisfied. The 46 states that were used consist of various phonemes and allophones.

The time spent in each state is the state occupancy, and can be modeled by various distributions (such as Gamma, Poisson and others). Speech parameters necessary for word reconstruction can be obtained analytically as well using the Hidden Semi Markov Model [2]. These parameters are language dependent. A discrete HSMM can be described by the following model :

A = $\{a_{ij}\}$ The state probability distribution

B = $\{b_j(k)\}$ The observational probability distribution

N The number of states in the model

D The state occupancy distributions where: $d_i(\tau)$ is the probability of staying in state *i* for τ time units

D = $\{d_1(\tau), d_2(\tau), \ldots, d_N(\tau)\}$

V The set of each state's possible observation symbols

V = $\{v_1, v_2, ..., v_n\}$

M The number of distinct observation symbols per state (size of VQ codebook)

where:

$\pi = \{\pi_i\}$ The initial state distribution

Thus, the HSMM is shown in notation as :

$$\lambda = \{\mathbf{A}, \mathbf{B}, \pi, \mathbf{D}\}$$

In Figure #9 we see a 5 state (N = 5) HSMM illustrating the possible state transition probabilities v(*i*, *j*) and the observational probability densities B(*i*, p) for each state *i* and next state *j*. Of note is the zero self-state transition probabilities v(*i*, *i*) for each state *i* = 1, ... , N.

4. SUMMARY

Confidence windowing techniques based on the most likely candidate for a missing phoneme can reconstruct TIMIT sentences when phonemes are removed. The method uses the window to locate all possible sequences in the sentences of the TIMIT database, and returns the most likely candidates for multiple removed phonemes.

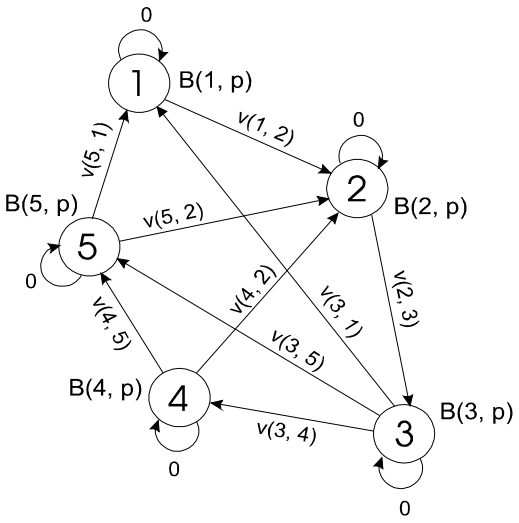


Figure #9. A hypothetical 5 state HSMM.

Current research includes the application of the Hidden Semi Markov Model (HSMM) for performing speech reconstruction. We are proving that the accurate results of the confidence window method indicate that additional probabilistic factors of the HSMM yield equal if not better results.

ACKNOWLEDGEMENT

This effort was sponsored by the Air Force Research Laboratory, Air Force Material Command, USAF, under agreement number F30602-00-1-0517. The U. S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotations thereon.

REFERENCES

[1] Scoville, Christopher W. *Spatially Dependent Probabilistic Events*, Master's Thesis, RPI, Troy NY, 1998.

[2] Nimal Ratnayake, *Phoneme Recognition Using a New Version of the Hidden Markov Model*, Ph.D. Thesis, RPI, Troy NY, 1992.

[3] Michael Moore, *Speech Reconstruction*, Internal RPI publication, RPI, Troy NY, October 2000.

[4] Michael Savic, *Speech Reconstruction*, Proposal to the Sponsor, RPI Troy NY, June 1996.