An Impact Matrix for the 1994 CSR Hub Evaluation

Ronald Rosenfeld

School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
roni@cmu.edu

The impact matrix on the following page summarizes the improvements achieved by the various sites on the unconstrained hub system (H1:P0) during this evaluation cycle, broken down into componental contributions. The numbers were compiled from measurements and estimates[1] provided voluntarily by the sites. The table includes all sites that responded to my survey. All numbers (except Word Error Rate) are reported as percent of WER improvement relative to some baseline.

The matrix is meant to serve the following purposes:

1. Present aggregate evidence to suggest that certain lines of inquiry have been fruitful while others have been less so.
2. Present a concise summary, *for each site independently*, of *some* of the componental improvements achieved by that site recently.
3. Allow a reader interested in a particular topic to quickly identify which system summary papers are likely to be relevant to that topic.

1. Caveats

The matrix is *not* meant to be used for cross-site comparisons of particular features or components. Such comparisons can be very misleading, for the following reasons:

- The baseline used to compute the relative improvements varies greatly among the sites, and often even within a site.
- Some numbers are based on adjudicated scores; others are based on pre-adjudicated scores.
- The reported improvements are based on different test sets (eval94, dev94, eval93, dev93, some combination, or some other sets), even within the same site.
- Sites reported only features that are new to them this year. Many of the features reported by some sites were implemented by other sites in previous years, and thus not included in their entry. Thus it is particularly important not to compare a filled entry with an empty one.
- Even within the same feature (row), there are wide variations in meaning. For example, "talker clustering" may mean simple gender-dependence for one site, or fancy 10-way talker classification for another.

Because of all of the above, no attempt was made to achieve self-consistency. In particular, the sum of all componental improvements reported for a specific site does not necessarily equal the total improvement achieved by that site during this last year on any test set.

2. Highlights

For some features, enough data points exist to draw the following conclusions:

- By far the biggest across-the-board improvement in WER this year came from increasing the vocabulary. WER reduction was generally proportional to the reduction in OOV rate of the test set. Several sites reported a saving of about 1.2 word recognition errors for every OOV token recovered.
- Adding acoustic training data had only a modest (though variable) impact. This is in spite of some sites' doubling the amount of training data to 120 hours!
- Adding language training data (whether more recent data, transcript data, or large amounts of data from dated test-set sources) had little effect. But lowering the ngram cutoffs did help some sites.
- Unsupervised acoustic speaker adaptation helped a variable amount.

[1] and, I should add, some wild guesses.

SITE	CU-HTK	IBM	LIMSI	AT&T	BBN	SRI	Dragon	Philips	BU	CMU	NYU	CU-CON
Primary P0 WER:	7.2	8.6	9.2	10.0	10.2	10.3	10.3	10.6	10.9	10.9	11.0	12.4
Acoustic Improvements:												
more training data		3	2	6	1		7		√			?
better signal processing							8					
talker clustering		1		8			12	2				?
wght optimization/insertion penalty	2								10			
quinphone+word boundary	5											
better state clustering					0		25	7	¬			
discrete→cont. HMMs					28							
better cross-word modeling						10				4		
unsprvsd speaker adaptation	10				2		1	5				
stochastic segment model									2			
segmental neural network					8							
Lexical Improvements:												
increased vocabulary	20	15	18	20	16	15	14	19	15	21		12
better phone set/dictionary	4											
better pronunciation models				11								
Language Improvements:												
more LM training data:												
LM setasides	¬	0	?		1			¬		?		
transcripts/VP data		1	?		1							
NYT/REU/WP data		0										
LM data cleanup			¬				¬					
LM conditioning fixup	2		2		¬		¬					
lower ngram cutoffs	¬	?		?	0		6	3		6		
4-gram	4	0			2							
5-gram				5								
class-based models		3			0							
maximum entropy models		¬										
LM caches								¬	1	3		
topic coherence model											4	
syntactic model											0	

Table 1: Impact Matrix for the 1994 CSR Hub Evaluation. *Do not compare figures across sites, or to empty entries!* See caveats on previous page.

Legend:

⟨*n*⟩ WER reduction of *n*% (relative) over some baseline.

¬ impact not measured, but presumed/known to be small ($\leq$ 2%).

√ impact not measured, but presumed/known to be large.

? impact not known.

(blank entry) feature not implemented, not reported, or already present in last year's system.

Send corrections, additions and updates to: `roni@cmu.edu`.