# DISSIMILARITY THREE WAYS

## BENGALI CORONALS SEEN THROUGH: ECHO REDUPLICATION CONFUSABILITY AND LEXICAL STATISTICS

SAMEER UD DOWLA KHAN, REED COLLEGE

IIT DELHI COLLOQUIUM, 18 NOVEMBER 2016

# GRADIENT SIMILARITY

- But some phenomena in perception and the lexicon are best described as involving gradient similarity
  - Lexical cooccurrence effects in Muna (Coetzee & Pater 2005)
    - is found in fewer roots with than with
  - Perceptual confusability in English (Cutler et al. 2004)
    - ! is misidentified as more often than as &
- It's possible that cases of supposed categorical identity/ similarity are in fact extreme cases of gradient similarity
  - cf. Vowel harmony in Hungarian (Hayes & Londe 2006)

- Echo reduplication
  - Most common in lgs across southern Asia, e.g. Hindi:
  - $ ' $ ‘name(s), etc.’
  - ( ' ( ‘bread, etc.’
- Phonological properties
  - Total reduplication
  - Systematic replacement of some material in reduplicant (RED) with one or more fixed segments
- Semantic properties
  - Typically denotes generalization: ‘X, etc.’, ‘superset of X’
  - In some lgs, it can also be disparaging

- Representative examples:

- Fixed C in Turkish (Southern 2005)
  - ‘box’ "

# ECHO REDUPLICATION

- Unlike prototypical reduplication, echo reduplication typically requires the base and RED to be non-identical
  - Unlike “emergence of the unmarked” cases of base-RED nonidentity, e.g. Sanskrit (Steriade 1988)
  - Unlike “default fixed segmentism”, e.g. Yoruba (Alderete et al. 1999)

- Presence of the fixed segment should be enough to generate base-RED nonidentity…
- …unless the fixed segment is identical to the segment it is meant to replace

# IDENTITY AVOIDANCE

- -initial words in Turkish have no echo form
  - $-$ ‘money’ " $-$ $-$ ‘money, etc.’
  - $ $ ‘table’ " * $ $ $ $ ‘towel, etc.’ " NO OUTPUT

- -initial words in Abkhaz take backup ! (Vaux 1996)
  - . & ‘fool’ " . & & ‘fool, etc.’
  - ‘money’ " * " ! ‘money, etc.’

- In Classical Tibetan , base takes backup (Beyer 1992)
  - . " . . ‘jumbled up’
  - . " . . ‘very stupid’
  - / " * / / " / / ‘miserable’

# IDENTITY AVOIDANCE

- Through various means, lgs work to avoid categorical identity between base and RED in echo forms
- Survey of echo forms in >100 lgs of India found identity avoidance in every case (Trivedi 1990)

- Previous work on echo forms generally describe a straightforward case of categorical identity avoidance
- No one has yet confirmed that this avoidance pattern does not extend to natural classes, or that it is not gradient

# A PUZZLE FROM ENGLISH

- What about English ! ?
  - !! $ 0 ‘doctor’ " $ 0 ! $ 0 ‘doctor$_{\text{DISMISSIVE}}$’
  - !! ‘school’ " ! ‘school$_{\text{DISMISSIVE}}$’
- Online survey, 190 respondents (Nevins & Vaux 2003)
- Identity avoidance: 95–97% of speakers rejected echo forms with ! for the 3 ! -initial words
  - !! ! ‘schmooze’ " * ! ! ‘schmooze$_{\text{DISMISSIVE}}$’
- Interestingly, 30% of speakers also rejected echo forms with ! for the one ! -initial word… why??
  - !! ! $ ‘schnozz’ " * ! $ ! $ ‘schnozz$_{\text{DISMISSIVE}}$’

# A PUZZLE FROM ENGLISH

- Possible explanations:

- The “two dialects” possibility
  - 65% of subjects obey identity avoidance
  - 30% obey categorical similarity avoidance, where ! and ! are of the same category: “sounds similar to ! ”

- The “matter of degree” possibility
  - 95% obey gradient similarity avoidance, of whom:
  - 65% considered ! and ! are sufficiently dissimilar
  - 30% considered ! and ! are excessively similar

# A PUZZLE FROM ENGLISH

- Another possible explanation: “this isn’t English”
  - Humorous and possibly peripheral to the language
  - Less common in English than in other lgs
  - ! is highly marked, restricted to borrowings from Yiddish
  - Construction is possibly borrowed from Yiddish (Southern 2005)

- To understand if echo reduplication can employ gradient

# QUESTIONS

- Does echo reduplication in Bengali involve…
  - Categorical identity avoidance,
  - Categorical similarity avoidance, or
  - Gradient similarity avoidance?
- If it is the latter, how can similarity be objectively measured on a gradient scale?
- As a comparison, we can investigate other parts of Bengali phonology that expected to employ this gradient similarity:
  - Lexical cooccurrence restrictions
  - Perceptual confusability

- Basic design:

- 60 stimulus words
  - Disyllabic stems
  - Content words: N, A, V (perfective participles)
- 2 registers of urban colloquial Bangladeshi Bengali
  - High register: closer to written Kolkata Standard
  - Low register: closer to eastern regional varieties
- Produced by adult female speaker
  - Proficient in both registers
  - 2 reps per variety = 240 recordings

# EXPERIMENT I: STIMULI

- 60 test words fell under 3 conditions:
- Identity: -initial words
- Similarity: words with -like initials
  - Coronal obstruents 1 3 3 4 41 !
- Control: words with non- -like initials
  - Coronal sonorants 6
  - Non-coronals 2

# EXPERIMENT I: STIMULI

- Consonants of Bangladeshi Standard Bengali (Khan 2010)

Identity Similarity Control

| | Labial | Dental | Alveolar | Post-Alv | Velar/Glot |
|---|---|---|---|---|---|
| Stop | 2 | 3 3 3 3 | 1 2 | | 1 . .2 |
| Affricate | | | 7 7 1 8 8 2 | | |
| Fricative | | | | ! | |
| Liquid | | 6 | | | |
| Nasal | | | | | (/) |

# EXPERIMENT I: SETUP

- 30 speakers of Bengali
  - Varied dialect background
  - Residents of CA
  - Paid $10
- Heard stimulus
  - Participant selected preferred register
  - Order randomized for each speaker
- Asked to produce echo reduplicated form
  - ‘cough’ "  ‘cough, etc.’ given as example
- Responses were transcribed

# EXPERIMENT I: HYPOTHESES

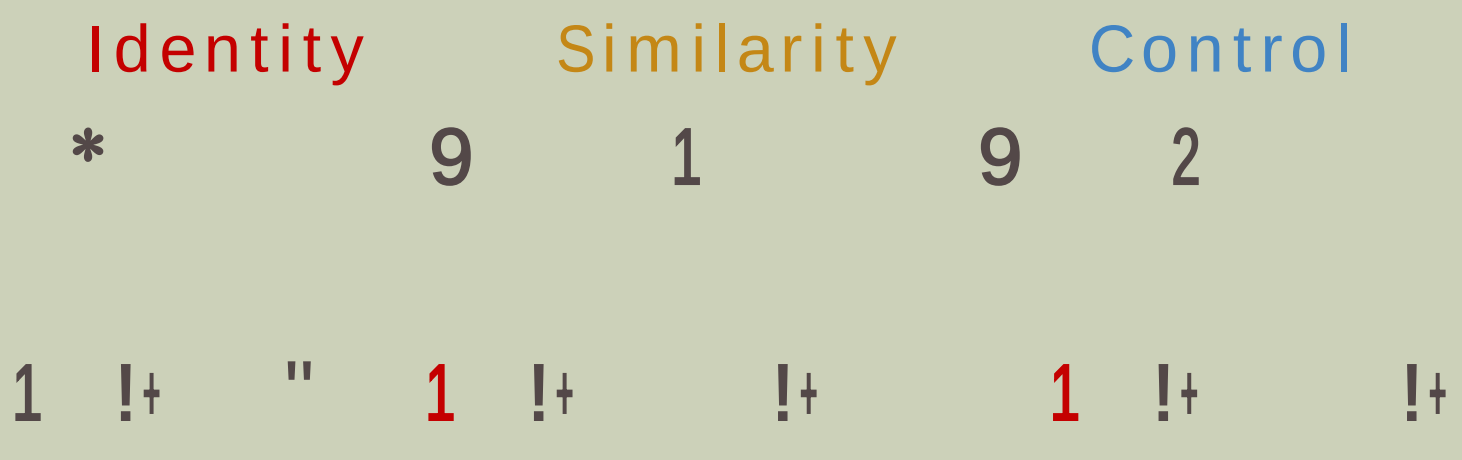
- Identity words will never use
- Control words will always use
- Similarity words are what are being tested:
  - Hypothesis 1: similarity = control (categorical identity)
  - Hypothesis 2: similarity = identity (categorical similarity)
  - Hypothesis 3: similarity is on a continuum

| Identity | Similarity | Control |
|---|---|---|
| * | 9 | 1 |
| | | 2 |

1 !+ " 1 !+ !+

# EXPERIMENT I: HYPOTHESES

- Identity words will never use
- Control words will always use
- Similarity words are what are being tested:
  - Hypothesis 1: similarity = control (categorical identity)
  - Hypothesis 2: similarity = identity (categorical similarity)
  - Hypothesis 3: similarity is on a continuum



# EXPERIMENT I: RESULTS

- Hypothesis 3 was borne out
- Similarity words lie on a continuum
  - Disprefer but not outright ungrammatical
  - Some consonants are more -like in behavior than others
- Seems like Cs that take less often are also phonetically closer to

# EXPERIMENT I: RESULTS

# EXPERIMENT I: DISCUSSION

- Echo reduplication in Bengali appears to incorporate a notion of gradient similarity avoidance
  - No straightforward clustering of consonants
  - Heavy overlap across clusters
  - Like the “matter of degree” hypothesis from English puzzle

# NEW QUESTIONS

- We should confirm our suspicion that our reduplication data can be modeled on an objective scale of similarity
- Is there a metric that Bengali speakers are using to calculate the similarity of an initial C and ?
- Metric has to be gradient, possibly language-specific

# SHARED NATURAL CLASSES

- Best-known option is shared natural classes (SNC) metric (Frisch et al. 1995/2004)
- Similarity of two Cs is based on the number of natural classes they share in the inventory
- Universal claim with language-specific application

- Hypothesis: the more natural classes shared between a C and , the less likely it will take in its echo RED

# SNC: METRIC

- In the SNC metric, similarity of $C_1$ and  is quantified as:

$$sim \quad \frac{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}{+}$$

- Compared SNC-similarity (line) to Exp 1 results (bars)

# SNC: DISCUSSION

- The SNC metric does an okay job overall $r^2$
- However, the area where it crucially fails to predict the data is the similarity set (coronal obstruents)

- The metric treats as inherently more similar to 3 and 4 than to 1 ... is there a way to adjust that?

# SNC: THOUGHT EXPERIMENT

- Original SNC metric derives directly from the phoneme inventory and feature set
- But what if we maintain the basic model but incorporate feature weights?

- Let's try a little thought experiment
- Weighting [dist] over [spread gl]: the - 3 distinction can be "heavier" than the - 1 distinction
- If this improves our metric, we can then pursue the question of whether these weights are justified

# WEIGHTED SNC: METRIC

- In an SNC-like model with feature weights, similarity of $C_1$ and is quantified as follows: (Wilson, p.c.)

$$sim \quad ! \quad i \quad !$$

$$f_i$$

!

- Where weights are drawn from the variation in the reduplication results, as follows:

# WEIGHTED SNC: METRIC

- Probability of use in the RED of a base with initial $C_1$

$$m \quad n \quad m"n \quad "sim \quad ^{n} \quad sim \quad m"n$$

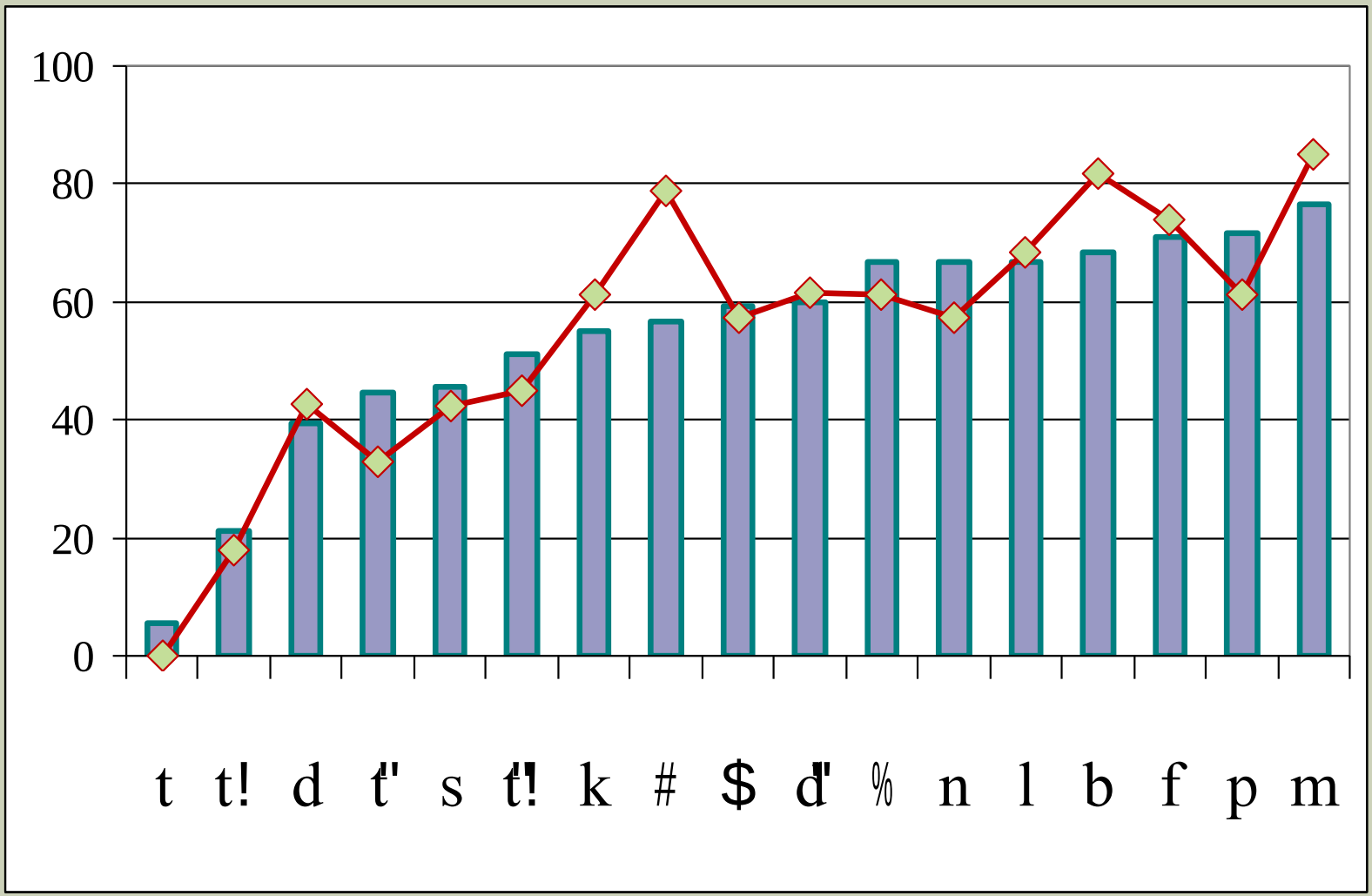
$$n$$

$$m$$

$$m$$

$$n$$

- Compared weighted similarity (line) to Exp 1 results (bars)


100
80
60
40
20
0
t
t!
d
t'
s
t"
k
#
$
d'
%
n
l
b
f
p
m

# WEIGHTED SNC: DISCUSSION

- With 4 adjusted feature weights, the SNC metric can closely model the reduplicative data $r^2$
  - [voice]: .554
  - [distributed]: .400
  - [strident]: .249
  - [spread glottis]: .198
  - All other features have a weight of 0.100

# NEW QUESTION

- Okay, but have we compromised the model?
- Is it no longer a similarity metric, but just a model of the reduplicative data?

- Let's see if our reduplicative data resemble other areas where gradient, lg-specific similarity is arguably relevant:
  - Lexical cooccurrence (McCarthy 1994)
  - Perceptual confusability (Shepard 1972)

- Similarity of two Cs is often negatively correlated with their cooccurrence within roots (Greenberg 1950)
  - English: two LAB or two DOR are underattested in [

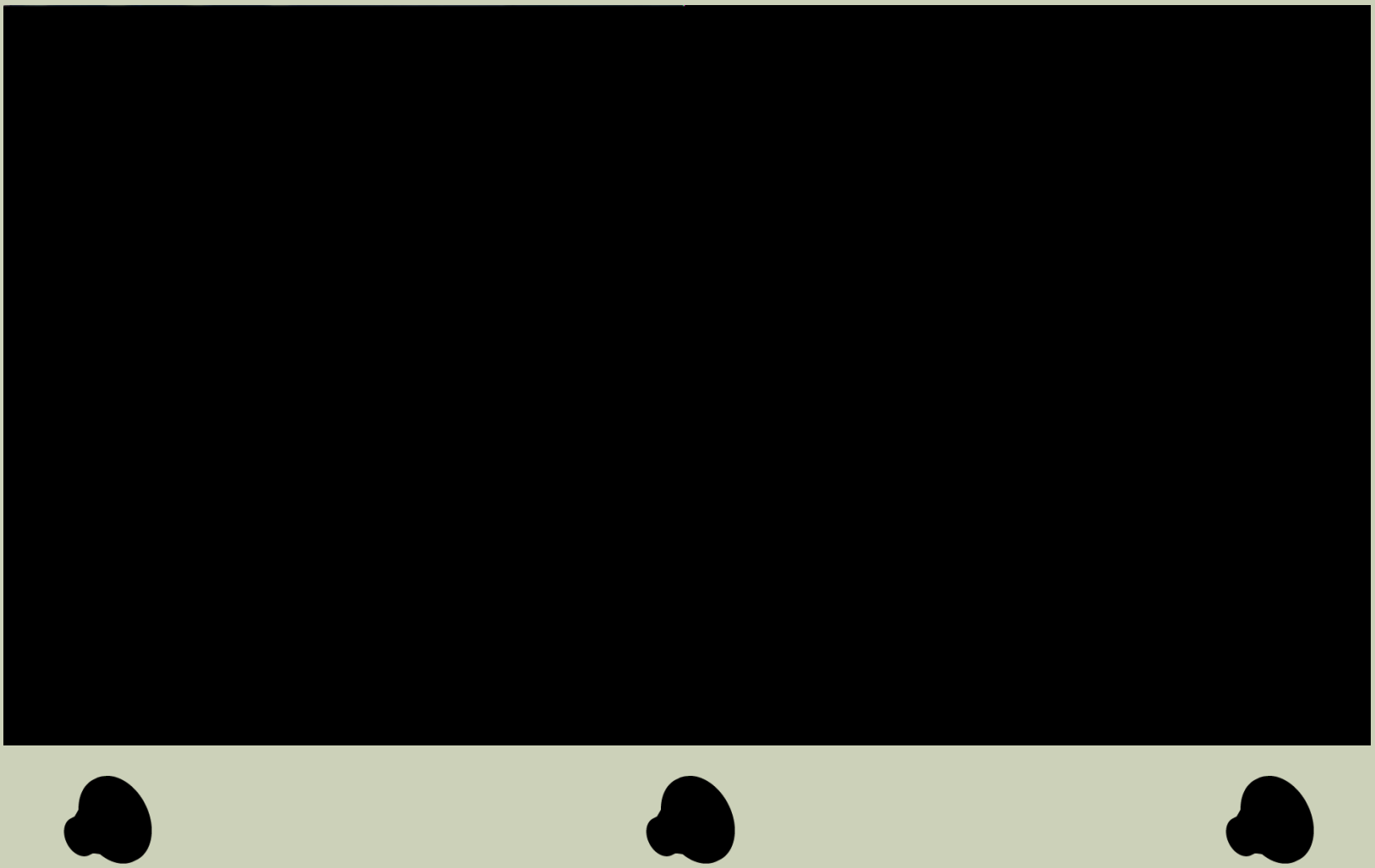
- Similarity of initial $C_1$ and medial  is the inverse of their observed / expected lexical

- The other area to look for the effects of gradient similarity is in perceptual confusability
  - Hindi: ( is misidentified as 3

# EXPERIMENT II: STIMULI

- 54 syllables
  - Onsets: 27 legal [Ca] syllables (all Cs but / ; )
  - Codas: 27 legal [aC] syllables (all Cs but 2 )
- Produced by adult female speaker
  - Best of several reps was normalized for amplitude
- Blocked by 3 masking conditions
  - Multi-talker babble
  - Pink noise
  - Quiet (no added sound)
- 54 syllables x 3 conditions x 3 reps = 486 trials

# EXPERIMENT II: TASK

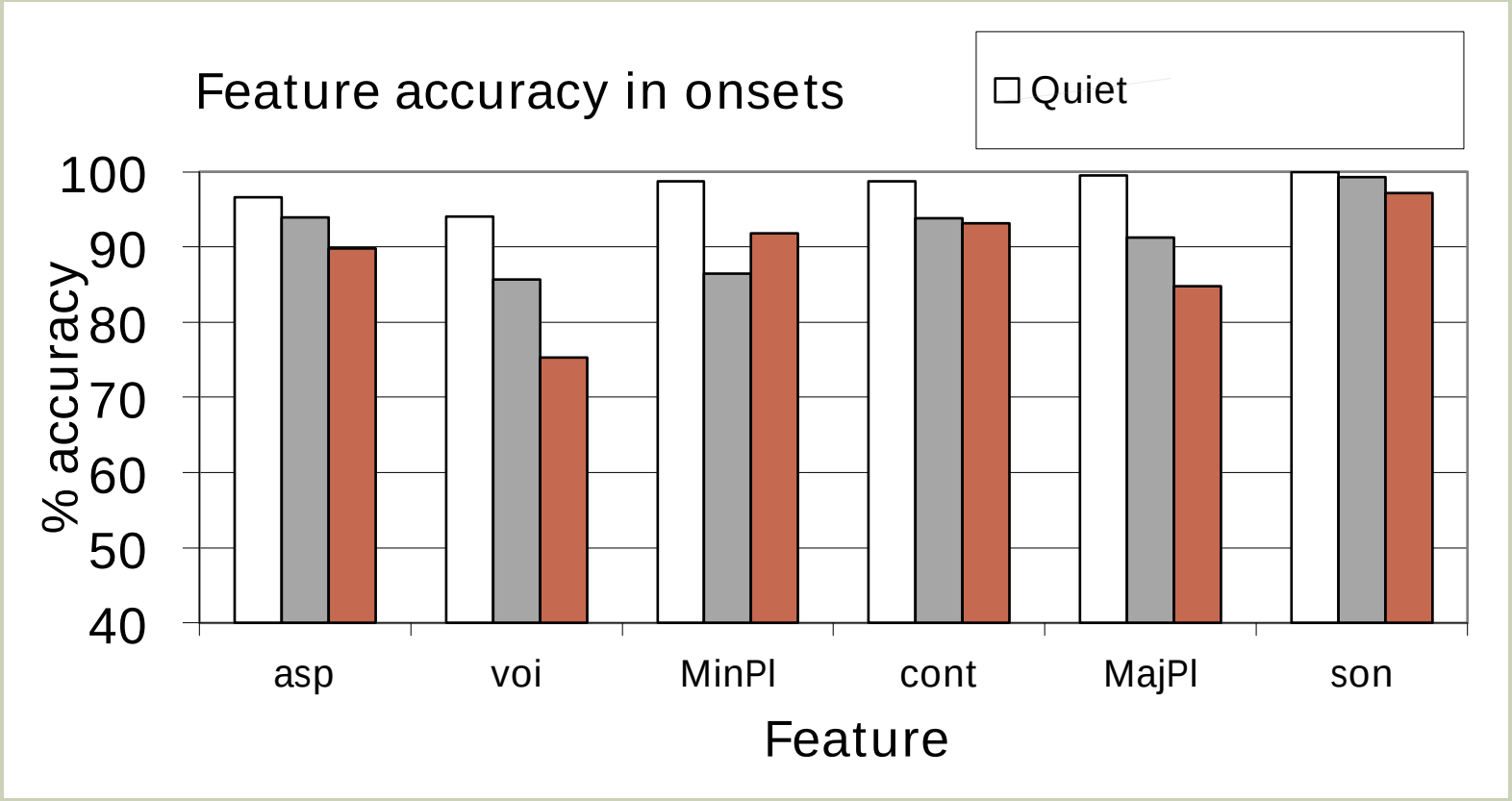


# EXPERIMENT II: HYPOTHESES

- The C most confused with  should be 1
  - Generalized: aspiration should be the most confusable feature
- Next most confused with  should be
  - Generalized: voicing should be the 2nd most confusable feature
- After that should be 3
  - Generalized: [distributed] and other minor place distinctions should be the 3rd most confusable
- After that should be
  - Generalized: [strident] and other manner-related distinctions should be less confusable than the preceding

# EXPERIMENT II: RESULTS

- Onset accuracy: 92% in quiet, 70% in noise, 60% in babble



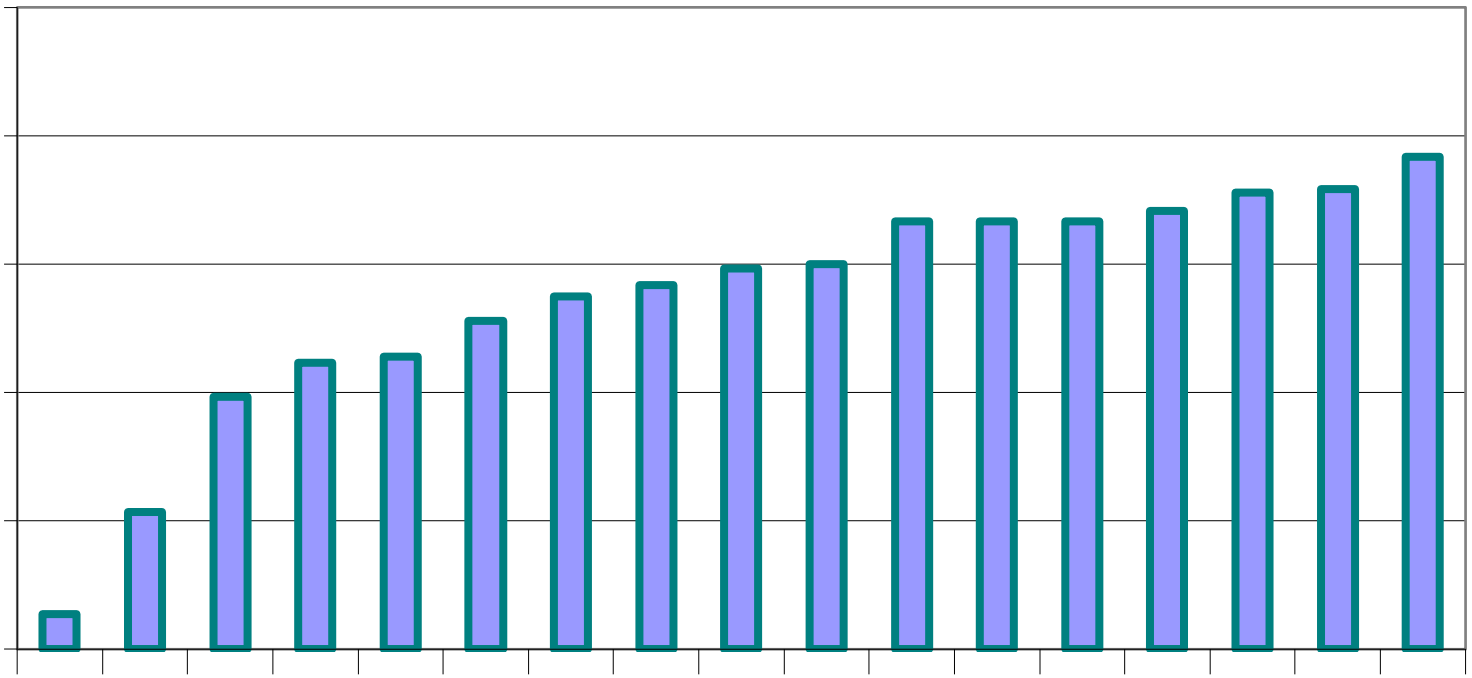
- Coda accuracy

# CONFUSABILITY: METRIC

- Similarity of $C_1$ and as drawn from confusion rate is quantified as follows: (Shepard 1972)

$$sim \qquad \frac{+}{+}$$

- Compared Exp 2 perceptions to Exp 1 productions
  - Removed “quiet” condition results (at ceiling)
  - Looked at onsets and codas separately



# CONFUSABILITY: DISCUSSION

- Consonant confusions with  in coda position are well correlated with the reduplicative results $r^2$

- But! Echo reduplication involves judging the similarity of onsets; why does the reduplicative data more closely resemble coda confusion?
  - Onset confusions with  were overall rare
  - Acoustic cues are perceptually less salient in codas (Wright 2004), so this is where similarity (not just identity) is likely more often relevant

- Okay, we need a recap.

- The current study demonstrates that fixed segment choice in Bengali echo reduplication is

# CONCLUSIONS

- The patterns clearly show that this similarity is gradient
- Many other phenomena previously treated as categorical have since been shown to be gradient
  - e.g. vowel harmony in Hungarian (Hayes & Londe 2006)

# CONCLUSIONS

- The current study proposes a modified version of the SNC metric of similarity
- I propose feature weighting for lg-specific application in diverse phonological phenomena

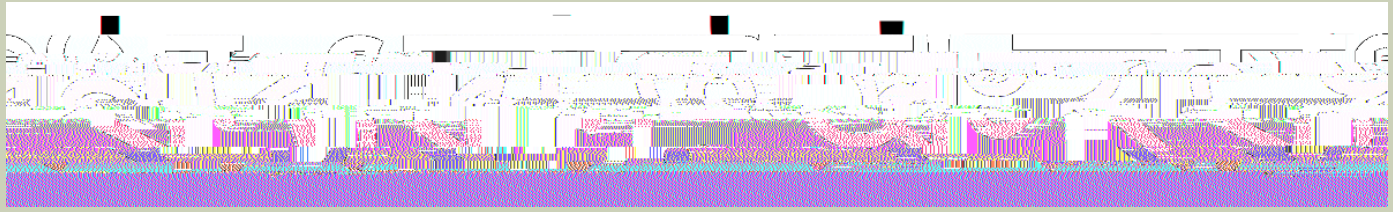
- The study also provides an interesting case in which the SNC metric can measure similarity in phonological phenomena *other than* lexical cooccurrence effects

# REMAINING QUESTIONS

- Is Bengali echo reduplication a special case, or should we look for gradient similarity in many more lgs?
- Why are the lexical cooccurrence effects of Bengali so different from the reduplicative results?
- How does this change as speakers deal with multiple phoneme inventories, e.g. bilinguals?

# ACKNOWLEDGMENTS


- This study was supported in part by Reed College's Stillman Drake Fund.
- To my participants and stimulus producers, to Colin Wilson (JHU), Kie Ross Zuraw (UCLA), Marc Garellek (UCSD), and Megha Sundara (UCLA), and to the audience at IIT Delhi:

[ o k o d on obad]

McCarthy, John J. 1994. The phonetics and phonology of Semitic pharyngeals. In Keating, Patricia (ed.) *Phonological structure and phonetic form: papers in laboratory phonology 3*, 191-233.