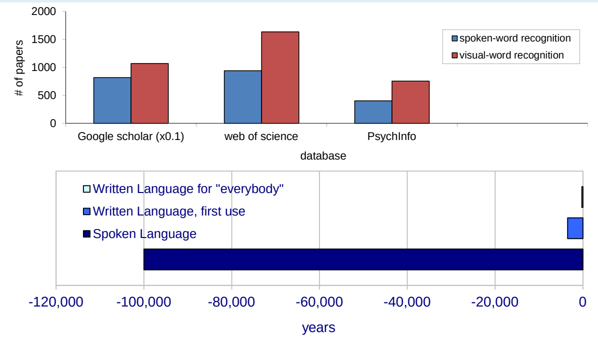
DADDY, EDDY, NINNY, NANNY and BALDEY: Big Data for speech perception

Anne Cutler

Radboud University Nijmegen
the MARCS institute
University of Western Sydney
CEDL
Max Planck Institute for Psycholinguistics

Research on spoken vs. visual language



Five data sets for speech perception research

Five data sets for speech perception research

- Speech perception research's scientific tradition: hypothesis-driven and experiment-based
- Big data of any kind notoriously hard to fund
- Often compiled by industry, or fully-funded government institutions
- Corpora: real life, undirected; but privacy issues.
- Who makes designed large data sets for speech perception research?

Five data sets for speech perception research

1. DADDY

Smits, R., Warner, N.L., McQueen, J.M. & Cutler, A. (2003). Unfolding of phonetic information over time: A database of Dutch diphone perception. *JASA*, **113**, 563-574.

http://www.mpi.nl/world/dcsp/diphones/index.html

(Sound files [both full and gated], plus all responses from 18 listeners)

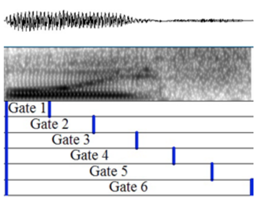
Why and how we collected this data set

Our Aim: Data to support a more realistic front end for a spoken-word recognition model, for all phonemes of a language, in all contexts where they could possibly occur.

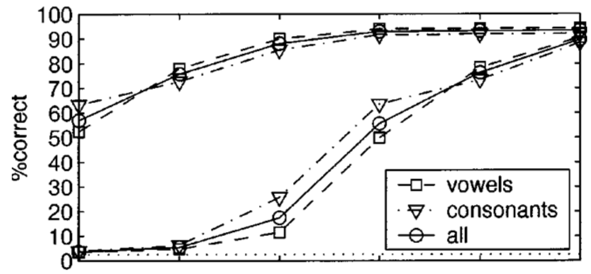
Experiment

- 2294 diphones: all possible within- or cross-word sequences of two Dutch phonemes including some stress variation (spoken by a single speaker)
- Each diphone gated to (mostly) 6 fragments (ending in square wave); Total = 13570 stimuli, randomised
- 18 listeners (judged phoneme 1 & 2)
- Total N responses per listener: 27140
- Average listener participation: 26 hrs
- Total database: 488520 data points

Incremental fragments

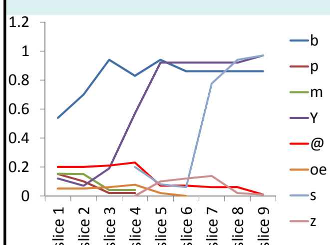


Orderly data!



% correct identifications for the diphones' Segment 1 (above), and segment 2 (below), across the 6 gated fragments of increasing size

DADDY data as front end for Shortlist B



The input *bus* presented to the model in 9 time slices; phoneme probabilities at each slice are computed from the DADDY data. These likelihoods continuously adjust the parameter P(E|W) which in turn adjusts P(W|E) i.e. the shortlist of lexical candidates.

likelihood — prior p

$$P(Word_i \mid Evidence) = \frac{P(Evidence \mid Word_i) \times P(Word_i)}{\sum_{j=1}^{j=n} P(Evidence \mid Word_j) \times P(Word_j)}$$

posterior p — normaliser

Five data sets for speech perception research

2. EDDY

Warner, N.L., McQueen, J.M. & Cutler, A. (2014). Tracking perception of the sounds of English. *JASA*, **135**, 2995-3006.

http://www.u.arizona.edu/~nwarner/WarnerMcQueenCutler.html

(Sound files and data files, for 20 listeners, as for DADDY)

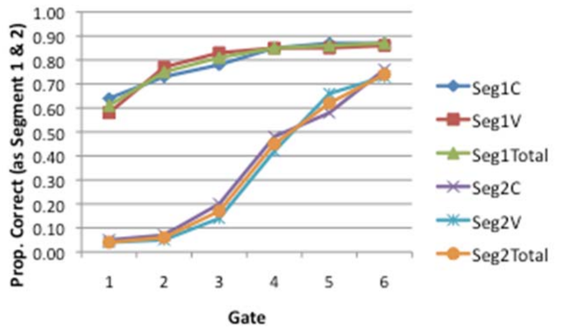
Why and how we collected this data set

Our Aim: Shortlist B works beautifully. An English front end would enable simulation of experiments in English, too.

Experiment

- All 2288 possible diphones of a variety of American English (spoken by a single speaker)
- Each diphone token again gated to (usually) 6 fragments (each ending in a square wave); Total: 13,464 stimuli
- 20 listeners judged all stimuli (1st and 2nd phoneme)
- Total number of responses per listener: 26928
- Average participation per listener: 33 one-hour sessions
- Total database: 538560 data points

More orderly data!



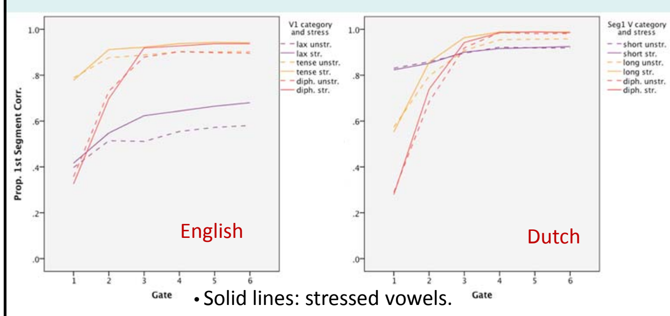
% correct identifications for the diphones' Segment 1 (above), and segment 2 (below), across the 6 gated fragments of increasing size

DADDY and EDDY can be compared, too

- Similar data sets, so: cross-language comparisons
- An example: stressed vs. unstressed vowels
- In Dutch, listeners attend to suprasegmental stress cues in recognising spoken words (e.g. *do-* from *DOminee* suffices to reject *domiNANT*)
- The same cues distinguish stressed from unstressed vowels in English, but English listeners rarely use them because inter-word distinctions rarely depend on it. (NB Dutch listeners to English do use the English cues!!)
- Are stress effects on vowel identification similar in the two languages?

(Cooper, Cutler & Wales, *Lg&Sp* 2002; Donselaar, Koster & Cutler, *QJEP* 2005; Cutler, *JASA* 2009)

Vowel identification in English & Dutch



- Solid lines: stressed vowels.
- Dashed lines: unstressed vowels.

Dutch: little stress effect (They are used to differently stressed vowels)
English: big effect (They don't expect vowels in multiple stress versions)

Five data sets for speech perception research

3. NINNY

Cutler, A., Weber, A., Smits, R. & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *JASA*, **116**, 3668-3678.

http://www.mpi.nl/people/cutler-anne/research

(Full identification response set from 16 native [American English] and 16 non-native [Dutch] listeners given American English CV or VC input)

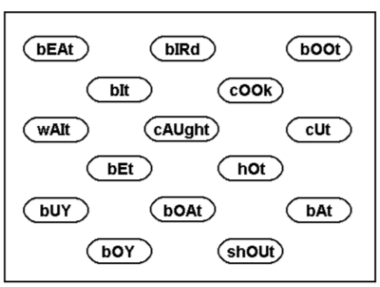
Why and how we collected this data set

Our Aim: Why exactly is non-native listening in noise so hard? If all predictability (lexical, any kind of contextual) is removed, do non-native listeners still suffer more from noise interference than native listeners? i.e. Do they always need better low-level evidence; or are they just less able to profit from higher-level predictability to recover from interference?

Experiment

- All possible CV and VC sequences of AmEng; 645 items
- In 3 levels of multi-talker babble noise (0, 8, 16 dB SNR)
- 32 listeners (16 each AmEng, Dutch) identified each phoneme of each syllable separately (3870 trials each)
- Total data set: 123840 data points

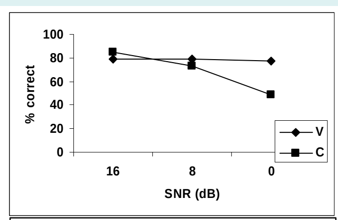
Response display



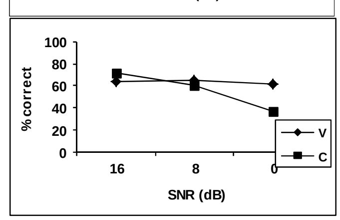
Separate displays for vowels, initial consonants and final consonants

Results

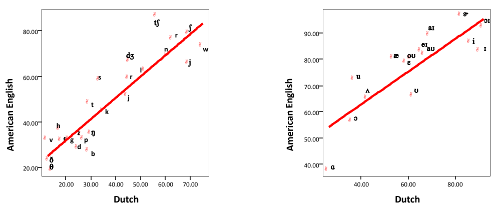
Native listeners (American English):



Non-native listeners (Dutch):



Results: Vowel and consonant identification



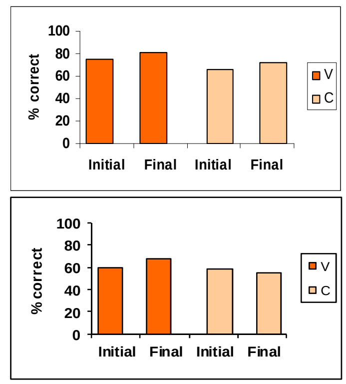
consonants vowels

Highly significant positive correlation (r = .91) between percent correct recognition per phoneme by native (vertical axis) and non-native listeners (horizontal axis)

Results: Position effects

Native listeners (American English):

Non-native listeners (Dutch):



Why is L2 listening in noise so hard?

- Noise masks non-native listening and native listening similarly
- The extra difficulty of non-native listening in noise is not due to phoneme identification problems alone
- It's because non-native listeners <u>can't recover from</u> these problems

Five data sets for speech perception research

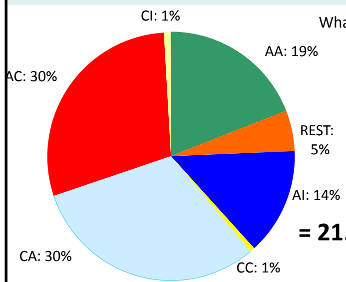
<u>4. NANNY</u>

Johnson, E.K., Lahey, M., Ernestus, M. & Cutler, A. (2013). A multimodal corpus of speech to infant and adult listeners. *JASA*, **134**, EL534-540.

Previously: Language input from 6 to 9 months

Van de Weijer (1999) *"Language Input for Word Discovery"*
3 months, all input heard by a single infant
3 weeks (85000 utterances) fully analysed

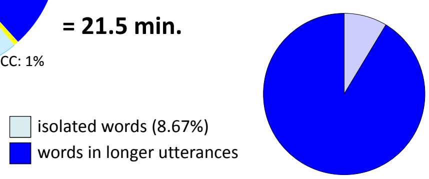
Previously: Language input from 6 to 9 months



What were the 85000 utterances?

1. Total Speaking Time: 2.56 hours per day
2. Speech types, by A(dult), C(hild) to A, C, I(nfant):

= 21.5 min.



(Van de Weijer, 1999)

→ Infants hear mostly continuous speech

Why and how we collected this data set

Our Aim: Answer some questions raised by existing corpora and provide relevant evidence on early word form acquisition.

Data Set

- 65 play sessions (33 hours of speech interaction) involving 28 triads, each of an 11-month-old infant with 2 caregivers
- Audio and (double) video record
- In part of the sessions, caregivers attempted to teach their infant new words
- In other parts, the caregivers interact with an experimenter and/or with each other or the infant

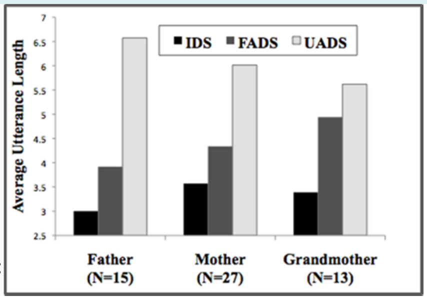
A word teaching example

The words were: a noun (e.g. *cactus*), a proper name (e.g. *Tigo*), a verb (e.g. *buigen* 'bow') and an adjective (e.g. *glanzend* 'shiny').

Double-view video allows eye gaze to be determined.

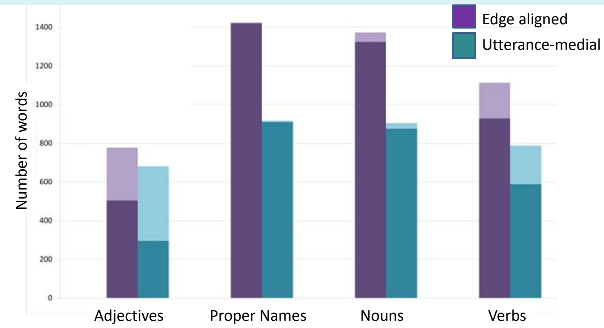
"Corpus" analysis: IDS vs. ADS

speaker:



Consistent with cross-corpus asymmetries, within this one corpus the difference between IDS and F(amiliar)ADS is much smaller than that between IDS and U(nfamiliar)ADS.

"Experimental" analysis: Word form segmentation



In agreement with the Edge Hypothesis, caregivers positioned target words at utterance edges. (Johnson, E.K., Seidl, A., Tyler, M.D. [2014]. The edge factor in early word segmentation: utterance-level prosody enables word form extraction by 6-month-olds. *PLoS ONE*, 9, e83546.)

Five data sets for speech perception research

5. BALDEY

Ernestus, M. & Cutler, A. BALDEY: A database of auditory lexical decisions. *Quarterly Journal of Experimental Psychology*, revision submitted, 2014.

http://www.mirjamernestus.nl/Ernestus/Baldey/index.html

(Sound files and Praat scripts for all 5541 items, and the full data set [accuracy, RTs] from 20 listeners)

Why and how we collected this data set

Our Aim: Data to support modelling of the lexical decision task and of recognition of spoken words of varying structure.

Well-understood task, but little data across types of words.

Experiment

- 5541 items; 2780 real Dutch words, 2761 pseudo-words
- 20 participants (10 M 10 F). 10 5-part sessions each.
- Realistic variation in word class (verb [regular, irregular], noun, adjective), length (1 to 5 syllables), morphology (stem+deriv 27.7%, stem+infl. 21.9%, stem+2 affixes 13.3%, simple 18.4%, compound 13.5%, compound+affix 5.2%)
- Pseudo-words (a) matched to real words on structural factors; (b) phonologically plausible
- 110420 timed responses

Lexical decision

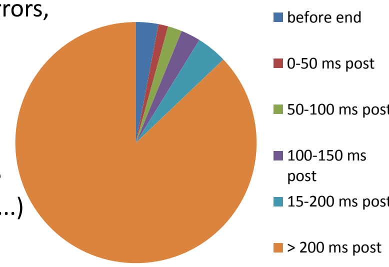
The nature of the lexical decision task

1. Words are heard in isolation. (So: no contextual support)
2. There are both words and non-words.

Thus to avoid making errors, listeners must be sure they have heard each entire stimulus item.

(even a beginning like *televisio-* might become a nonword with *-d* or *-z*...)

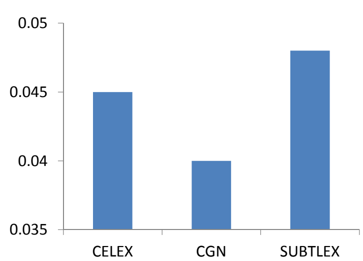
Our data show that our listeners performed the task appropriately.



Comparing corpora via this data set!

Data set offers many analysis options.
We include frequency measures from several corpora: CELEX, Corpus of Spoken Dutch (CGN), SUBTLEX.

Averaging across all word types, correlation of log RT measured from word offset with log word-form frequency in each of these corpora:



Five data sets for speech perception research

- Speech perception research's scientific tradition: hypothesis-driven and experiment-based
- Big experimental data sets allow testing of many hypotheses beyond those that motivated them
- Over to you....