Precision of Phoneme Boundaries Derived using Hidden Markov Models

Greg Kochanski, Ladan Baghai-Ravary, and John Coleman

University of Oxford Phonetics Laboratory, UK

Supported via ESRC RES-062-23-1172.

Presented at Interspeech 2009, 7-10 September 2009, Brighton UK. Talk at http://kochanski.org/gpk/papers/2009/LadanInterSpeech . Based on the paper at http://kochanski.org/gpk/papers/2009/IS090244.pdf : Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009). ISSN 1990-9772 Brighton, UK, 7-10 September 2009. Pp. 2879-2882.

Precision of Phoneme Boundaries Derived using Hidden Markov Models

--or--

Do different forced alignment systems agree? If so, on what types of boundaries?

Ladan Baghai-Ravary, Greg Kochanski, and John Coleman

University of Oxford Phonetics Laboratory, UK

Supported via ESRC RES-062-23-1172.

Why do we care?

Statistics

F-tests, chi-squared tests, and ANOVA typically assume that the variances of data are equal.
If not, the data must be weighted according to its variance when evaluating alignment systems.

Curiosity

Humans have more trouble with some borders than others. Do machines have difficulties on the same phonemes? If not, what causes the difference?

Experimental Design

In Phonetics/Psychology experiments, you may want to design the experiment to only use borders that can be precisely defined.

Speech Technology

Synthesizers glue fragments of speech together and high quality output requires that the fragments be cut consistently. Perhaps one can avoid the difficult borders?

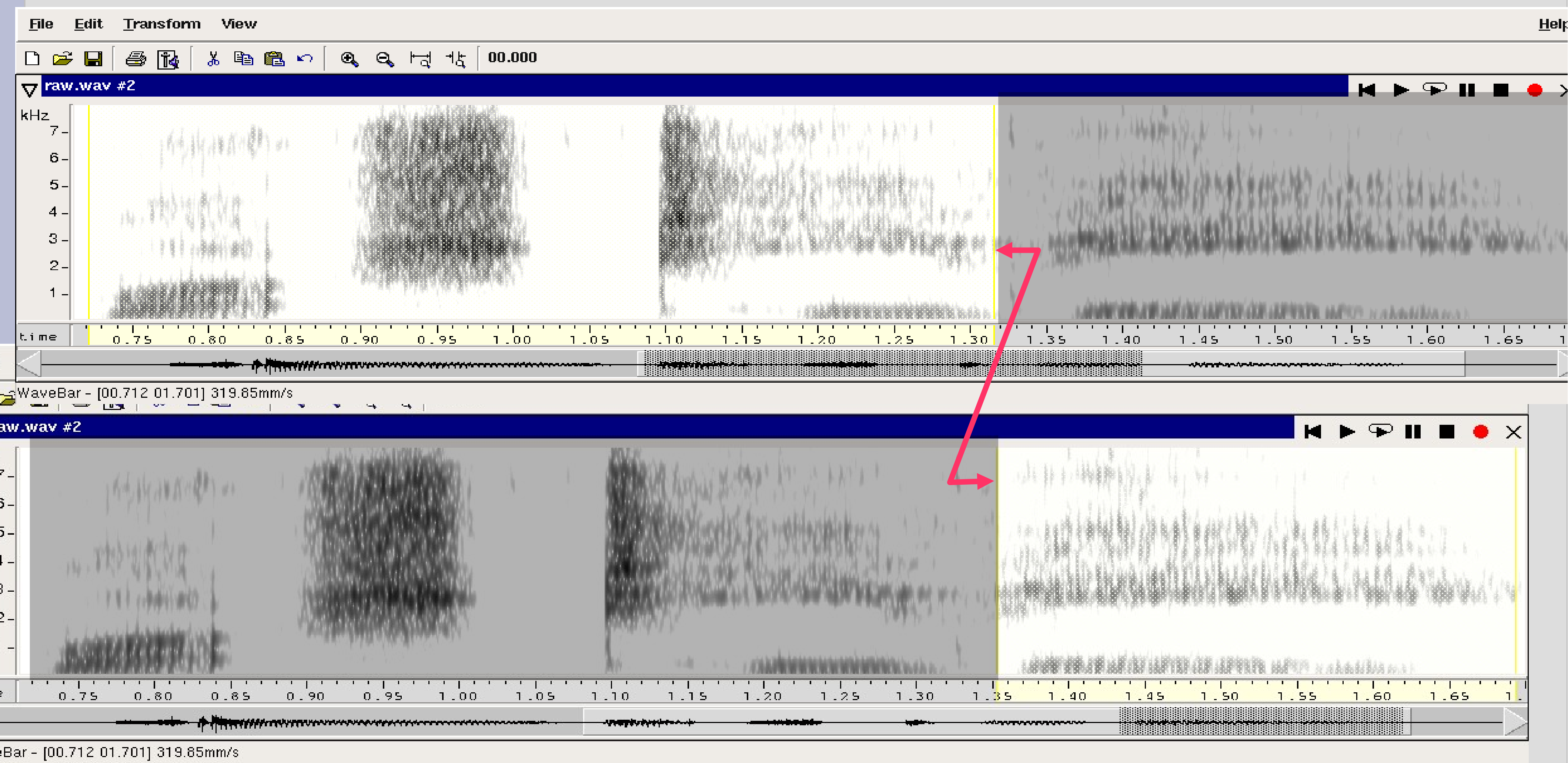
Why not measure against a human "gold standard"?

- Human segmentation is expensive: ~US$10,000 per hour of speech
- What's so good about human segmentation anyway?
  - Humans do not normally locate segment boundaries
    - Identifying phonemes may be a natural task
    - But, locating borders requires training
  - Oddly, phoneme boundaries are normally placed visually.
  - Serious human segmentation projects always have written rules
    - This suggests that rules are necessary
      - Without rules, humans some boundaries are ambiguous
    - Presumably, these rules affect the segmentation
      - So, human segmentation isn't entirely natural
      - Somewhat arbitrary.

What to measure?

Ask the application:

- Concatenative speech synthesis (cutting and glueing)

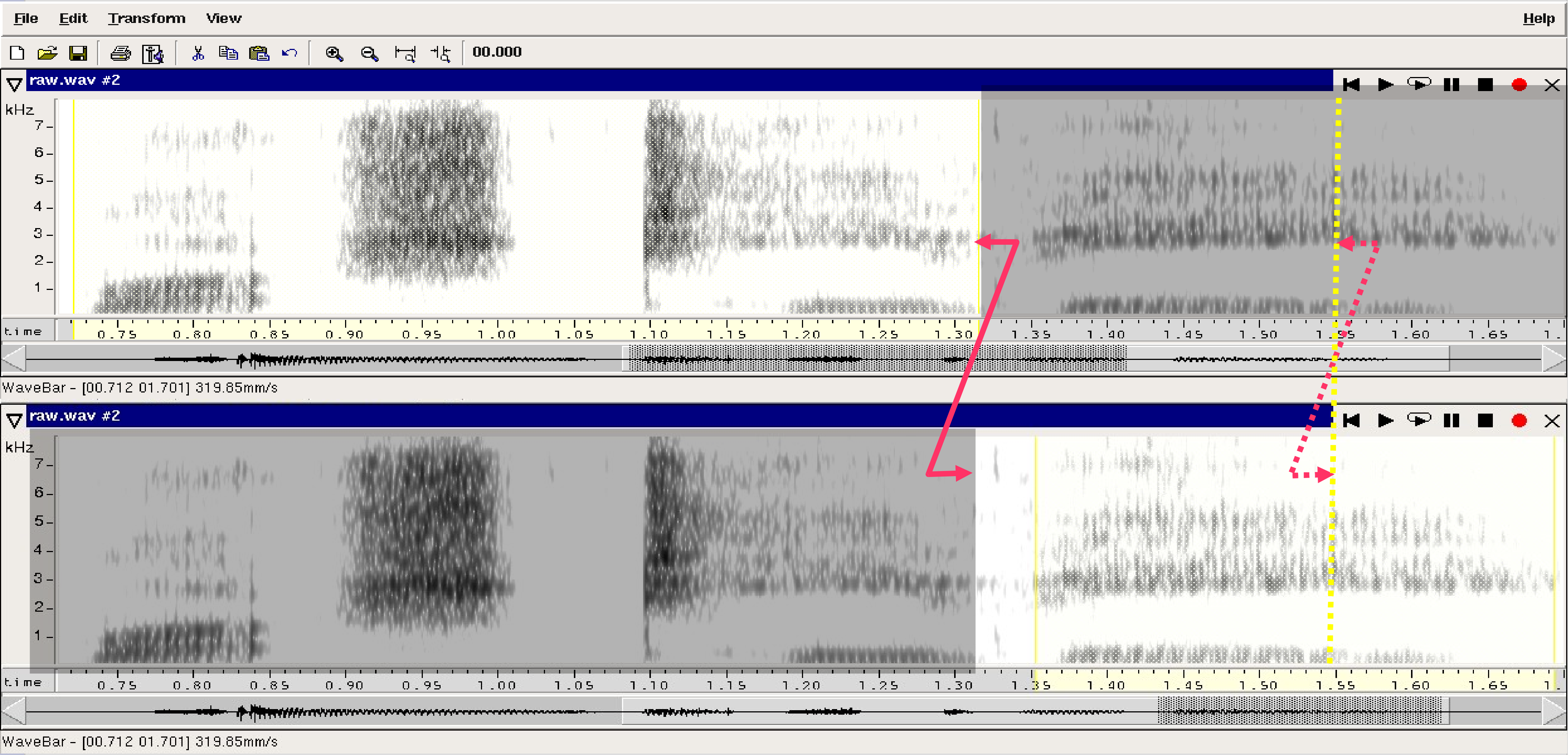


Consistency matters

What to measure?

Ask the application:

- Concatenative speech synthesis (e.g. cutting and glueing of diphones or larger units)



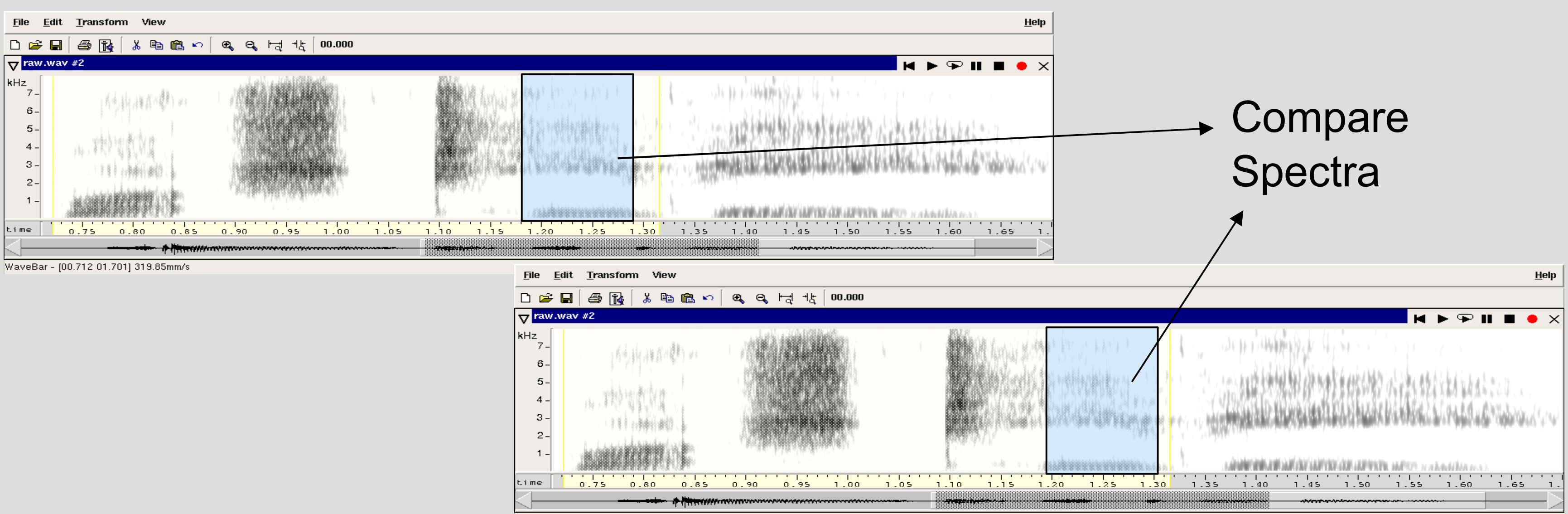
The location of cuts doesn't matter (much) as long as they are consistent.

What to measure?

Ask the application: Phonetics – comparing two dialects

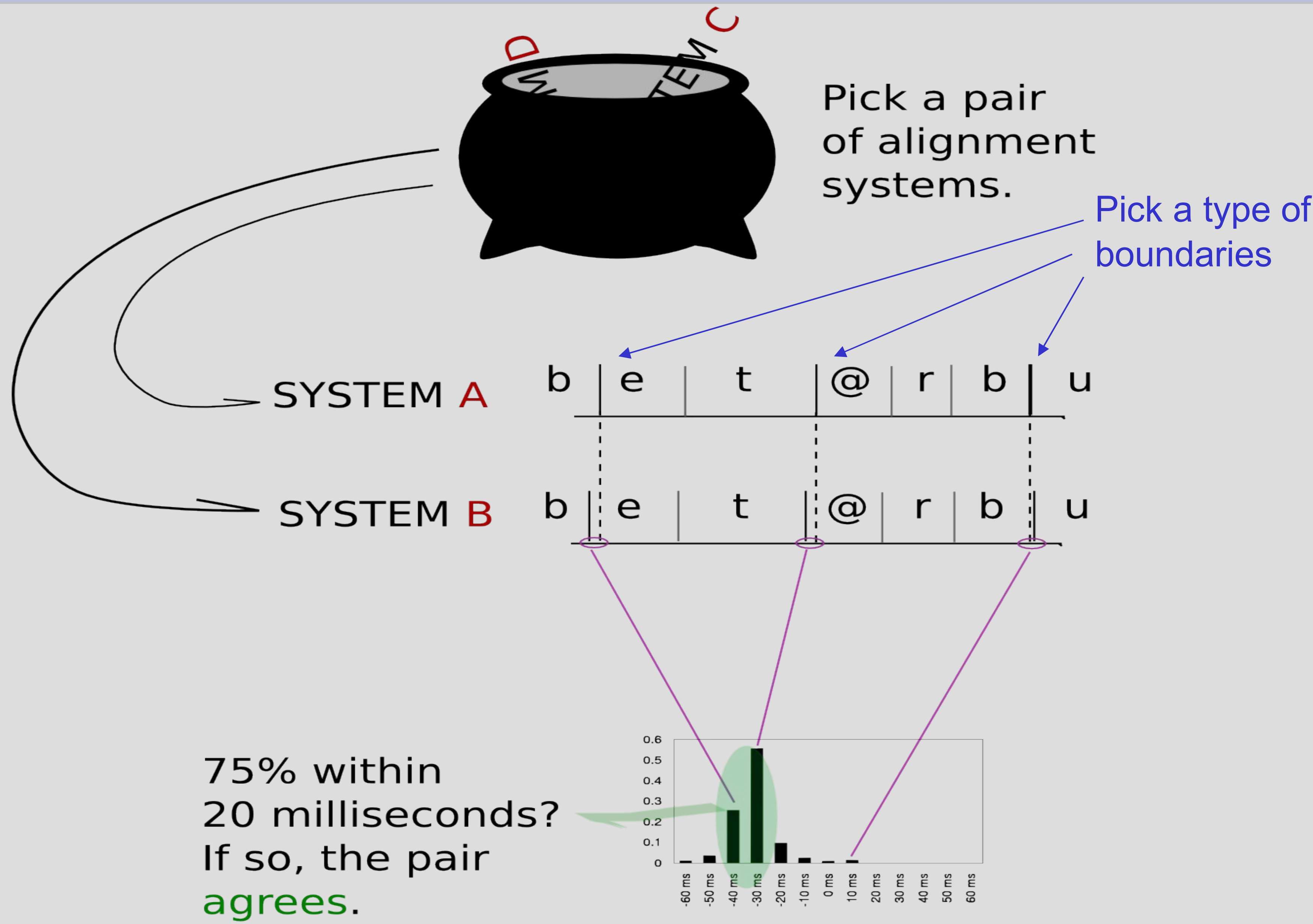
- Cut speech into phoneme regions.
- Measure acoustic properties within regions.

- Regions in the two dialects absolutely must be consistent.
- Regions need to be matched to human definitions only if you want to write about some particular phoneme.



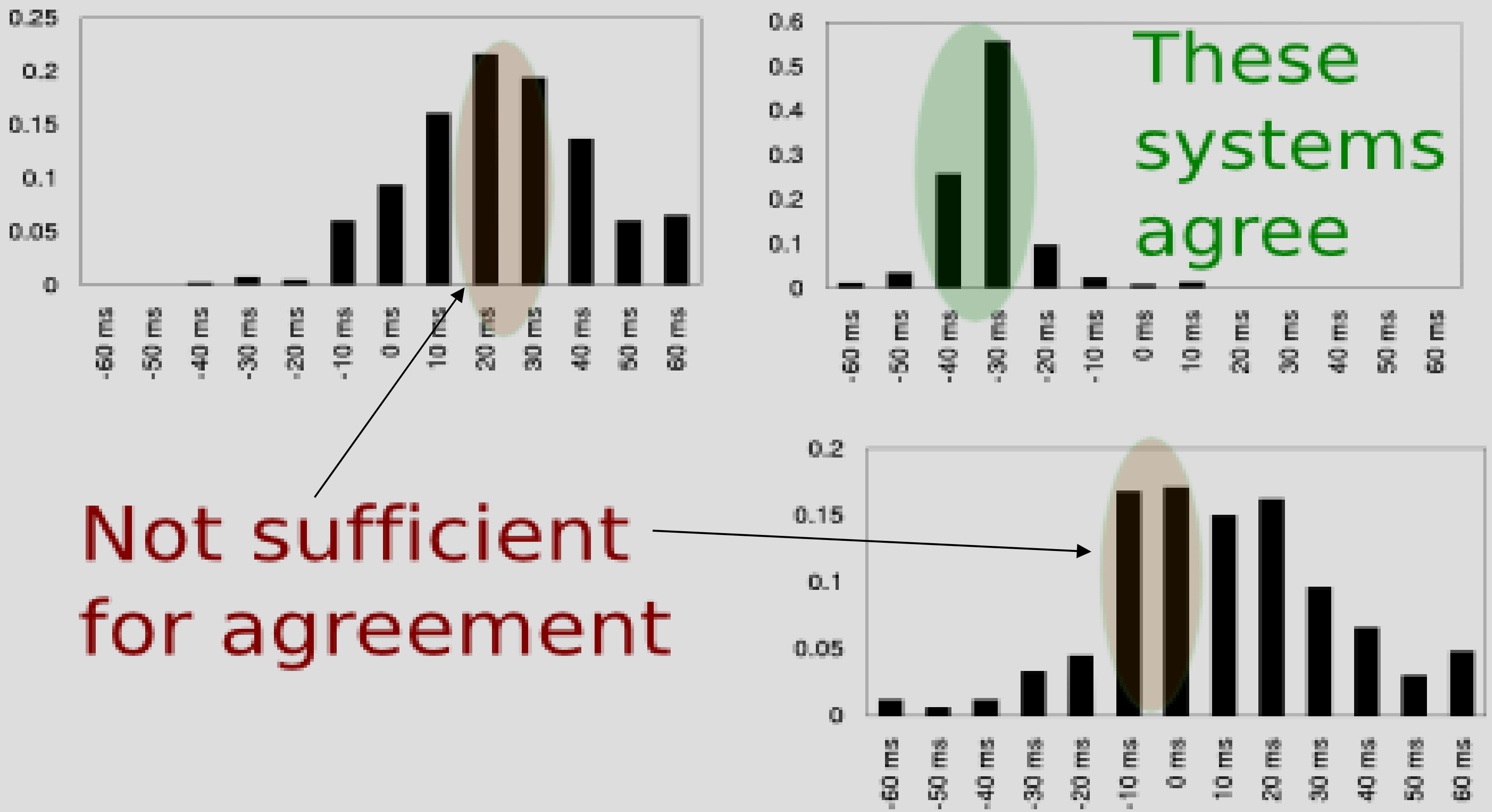
The location of cuts don't matter (much) as long as they are consistent.

How to measure?



Repeat for all pairs; count the number of agreements.

Algorithm



...then count the number of systems that agree for that particular transition.

Training Data and Technologies

Training Data:

Ad hoc, simple read sentences and citation form words.

23,000 utterances, 48,000 total words, vocabulary size=16,000.

16kHz sampling rate.

Technology:

HTK (Cambridge UK) toolkit, monophone models.

“short pause” phoneme inserted between words.

Three groups of systems:

* MFCC front end, 20ms window
* LP-Cepstra front end, 24ms window.
* Auditory description vectors front end
  - Cube-rooted erb-wide filters
  - Edge detectors
  - Voicing detector, etc
  - Reduced via principle components analysis from 51 dimensions to 19

48 alignment systems in total.

HMM details:

1, 2, 4, or 8 Gaussians per state

2 or 3 states per phone (strict left-to-right)

4 or 5 states per phone (allowing skips).

Results

Class	Abbr.	British English Phonemes (SAMPA)
Nasal	Nas	m, n, N
Plosive	Plo	b, d, g, k, p, t
Affricate	Aff	dZ, tS
Fricative	Frc	D, S, T, Z, f, h, s, v, z
Vowel	Vow	@, A, E, I, O, U, V, {, 3, i, u, Q
Approximant	App	r, j, l, w
Diphthong	Dip	I@, U@, aI, aU, E@, eI, OI, @U
Silence	Sil	silence, short inter-word pause

Classes of phonemes:

Compute the histograms for all transitions from one class to another.

Why classes? Some phoneme-phoneme transitions are quite rare.

Some transitions have vastly more agreement than others.

Number of agreements out of 1128 possible pairs.

Transitions Class A → Class B		Class B							
		Plo	Aff	Frc	Nas	App	Vow	Dip	Sil
Class A	Plo	19	39	70	84	166	411	469	57
	Aff	369	–	25	–	63	688	753	123
	Frc	438	86	22	456	133	423	537	116
	Nas	241	632	247	5	52	332	396	84
	App	208	156	251	60	5	26	9	103
	Vow	332	365	254	182	8	1	1	78
	Dip	361	437	260	170	6	0	4	77
	Sil	923	395	299	817	713	513	516	–

Results

Class	Abbr.	British English Phonemes (SAMPA)
Nasal	Nas	m, n, N
Plosive	Plo	b, d, g, k, p, t
Affricate	Aff	dZ, tS
Fricative	Frc	D, S, T, Z, f, h, s, v, z
Vowel	Vow	@, A, E, I, O, U, V, {, 3, i, u, Q
Approximant	App	r, j, l, w
Diphthong	Dip	I@, U@, aI, aU, E@, eI, OI, @U
Silence	Sil	silence, short inter-word pause

Transitions Class A → Class B		Class B							
		Plo	Aff	Frc	Nas	App	Vow	Dip	Sil
Class A	Plo	19	39	70	84	166	411	469	57
	Aff	369	–	25	–	63	688	753	123
	Frc	438	86	22	456	133	423	537	116
	Nas	241	632	247	5	52	332	396	84
	App	208	156	251	60	5	26	9	103
	Vow	332	365	254	182	8	1	1	78
	Dip	361	437	260	170	6	0	4	77
	Sil	923	395	299	817	713	513	516	–

* Transitions between similar phonemes are very uncertain

Results

Class	Abbr.	British English Phonemes (SAMPA)
Nasal	Nas	m, n, N
Plosive	Plo	b, d, g, k, p, t
Affricate	Aff	dZ, tS
Fricative	Frc	D, S, T, Z, f, h, s, v, z
Vowel	Vow	@, A, E, I, O, U, V, {, 3, i, u, Q
Approximant	App	r, j, l, w
Diphthong	Dip	I@, U@, aI, aU, E@, eI, OI, @U
Silence	Sil	silence, short inter-word pause

Transitions Class A → Class B		Class B							
		Plo	Aff	Frc	Nas	App	Vow	Dip	Sil
Class A	Plo	19	39	70	84	166	411	469	57
	Aff	369	–	25	–	63	688	753	123
	Frc	438	86	22	456	133	423	537	116
	Nas	241	632	247	5	52	332	396	84
	App	208	156	251	60	5	26	9	103
	Vow	332	365	254	182	8	1	1	78
	Dip	361	437	260	170	6	0	4	77
	Sil	923	395	299	817	713	513	516	–

* The table is somewhat symmetrical (ie A|B agrees if B|A agrees) but with some dramatic exceptions: pairs involving silence and plosives.

Why silence? Sentences start sharply and end gradually.
Why plosives? The beginning of a plosive is much different from the end.

Inside the classes: a look at individual phonemes

The fewest agreements (<5% agree):

/k/ or /v/ → Silence

/aI/ or /aU/ → /@/

…

Most agreements (>99% agree):

Silence → /S/, /d/, /b/, …

/s/ → /@U/, /eI/, /aI/, /O/, …

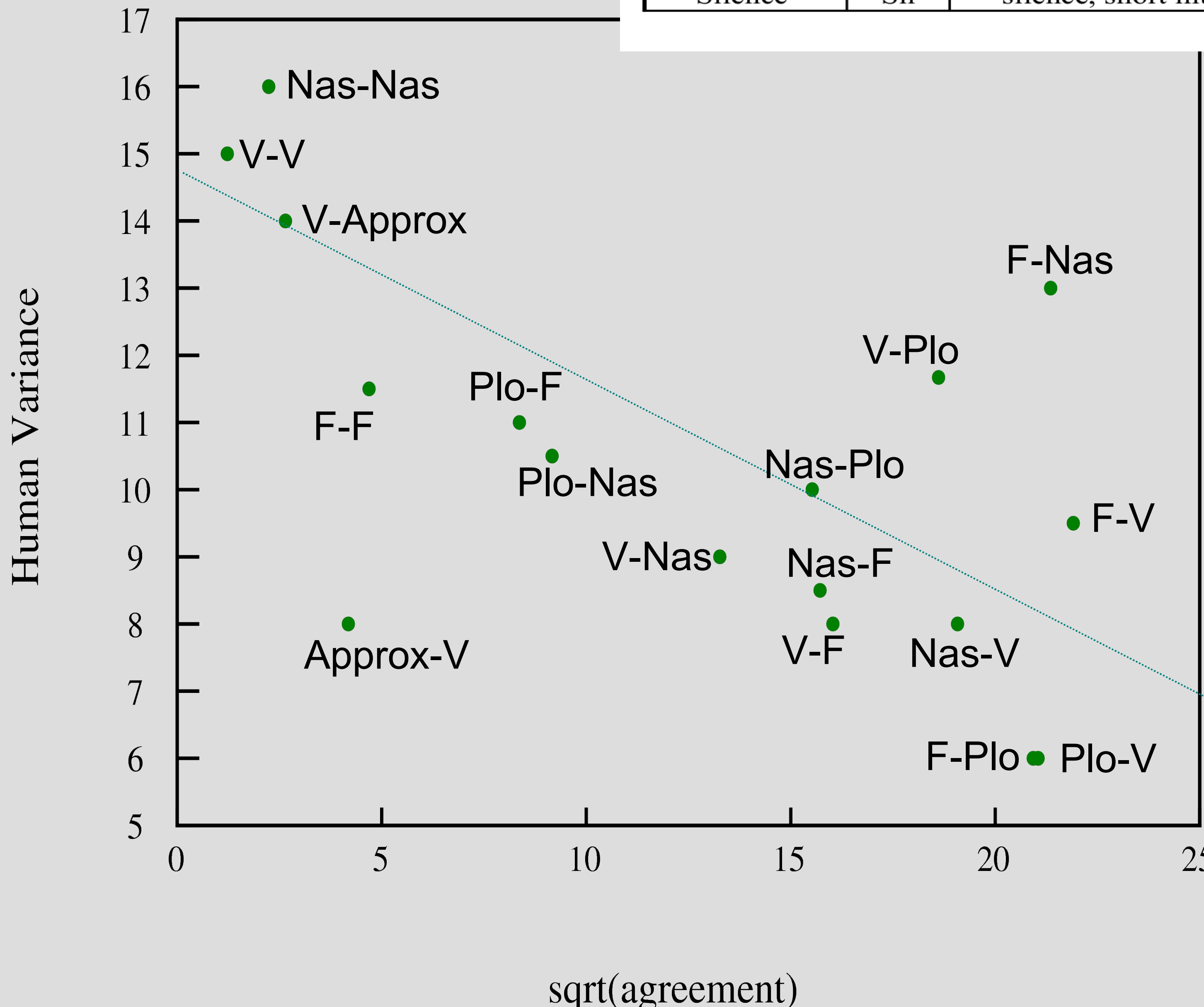
...

Transitions Class A → Class B		Class B							
		Plo	Aff	Frc	Nas	App	Vow	Dip	Sil
Class A	Plo	19	39	70	84	166	411	469	57
	Aff	369	–	25	–	63	688	753	123
	Frc	438	86	22	456	133	423	537	116
	Nas	241	632	247	5	52	332	396	84
	App	208	156	251	60	5	26	9	103
	Vow	332	365	254	182	8	1	1	78
	Dip	361	437	260	170	6	0	4	77
	Sil	923	395	299	817	713	513	516	–

Individual phoneme results are generally consistent with class-based results.

Comparison to human labelling

Class	Abbr.	British English Phonemes (SAMPA)
Nasal	Nas	m, n, N
Plosive	Plo	b, d, g, k, p, t
Affricate	Aff	dZ, tS
Fricative	Frc	D, S, T, Z, f, h, s, v, z
Vowel	Vow	@, A, E, I, O, U, V, {, 3, i, u, Q
Approximant	App	r, j, l, w
Diphthong	Dip	I@, U@, aI, aU, E@, eI, OI, @U
Silence	Sil	silence, short inter-word pause



Vertical: standard deviation of human labellers, from M.B. Wesenick and A. Kipp 1996.

Horizontal: square root of our agreement scores.

Our side: vowels lumped with dipthongs.

Their side: voiced and unvoiced plosives lumped, voiced and unvoiced fricatives lumped.

Conclusions

- Substantial differences between different boundaries
- Some correlation with human errors.
- Precision tends to be lower for similar phonemes
- Precision best at beginnings,
- Worst at ends.
- When evaluating alignment systems, weight the boundaries.

Why do we care?

Statistics

F-tests, chi-squared tests, and ANOVA typically assume that the variances of data are equal.
If not, the data must be weighted according to its variance when evaluating alignment systems.

Curiosity

Humans have more trouble with some borders than others. Do machines have difficulties on the same phonemes? If not, what causes the difference?

Experimental Design

In Phonetics/Psychology experiments, you may want to design the experiment to only use borders that can be precisely defined.

Speech Technology

Synthesizers glue fragments of speech together and high quality output requires that the fragments be cut consistently. Perhaps one can avoid the difficult borders?

Why not measure against a human “gold standard”?

- Human segmentation is expensive: ~US$10,000 per hour of speech

- What's so good about human segmentation anyway?
  - Humans do not normally locate segment boundaries
    - Identifying phonemes may be a natural task
    - But, locating borders requires training

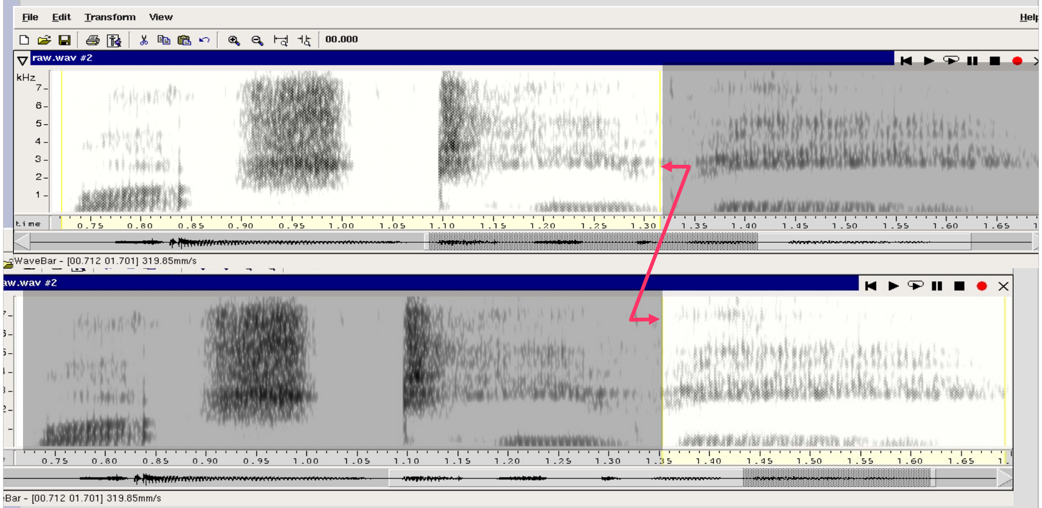
  - Oddly, phoneme boundaries are normally placed visually.

  - Serious human segmentation projects always have written rules
    - This suggests that rules are necessary
      - Without rules, humans some boundaries are ambiguous
    - Presumably, these rules affect the segmentation
      - So, human segmentation isn't entirely natural
      - Somewhat arbitrary.

What to measure?

Ask the application:

- Concatenative speech synthesis (cutting and glueing)

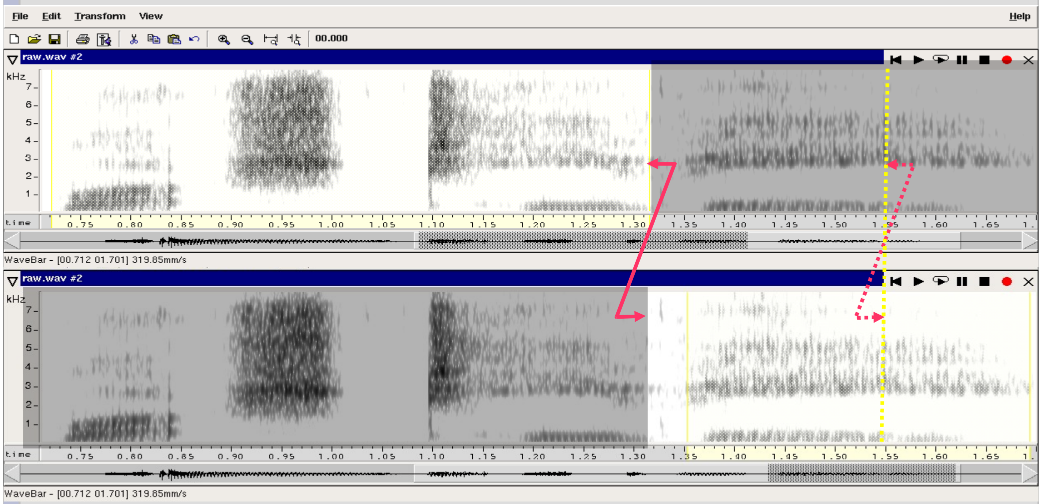


Consistency matters

What to measure?

Ask the application:

- Concatenative speech synthesis (e.g. cutting and glueing of diphones or larger units)



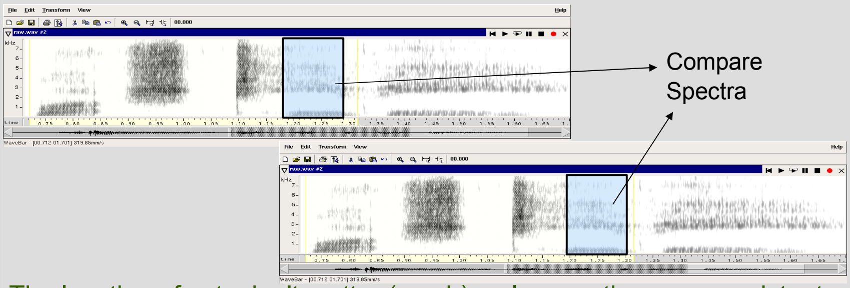
The location of cuts doesn't matter (much) as long as they are consistent.

What to measure?

Ask the application: Phonetics – comparing two dialects

- Cut speech into phoneme regions.
- Measure acoustic properties within regions.

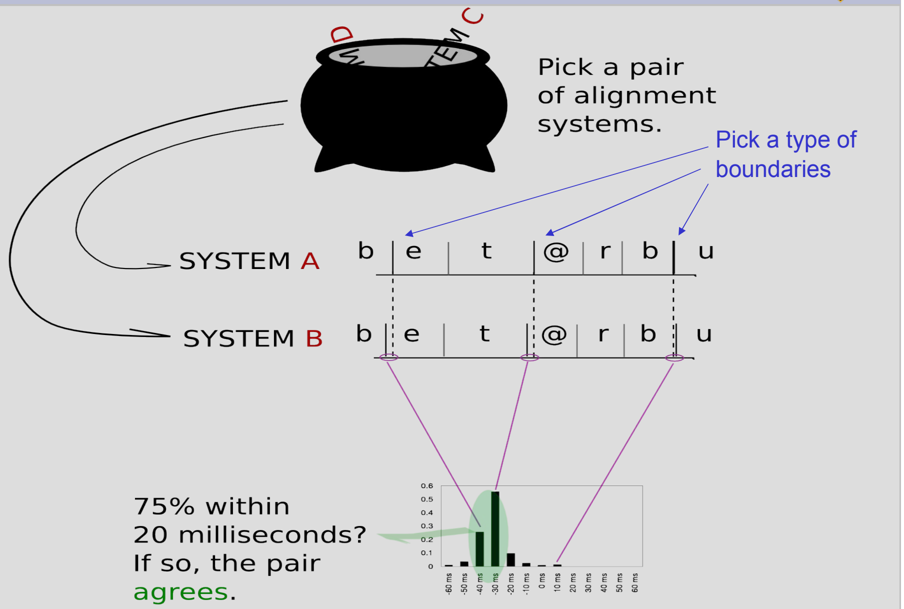
- Regions in the two dialects absolutely must be consistent.
- Regions need to be matched to human definitions only if you want to write about some particular phoneme.



Compare Spectra

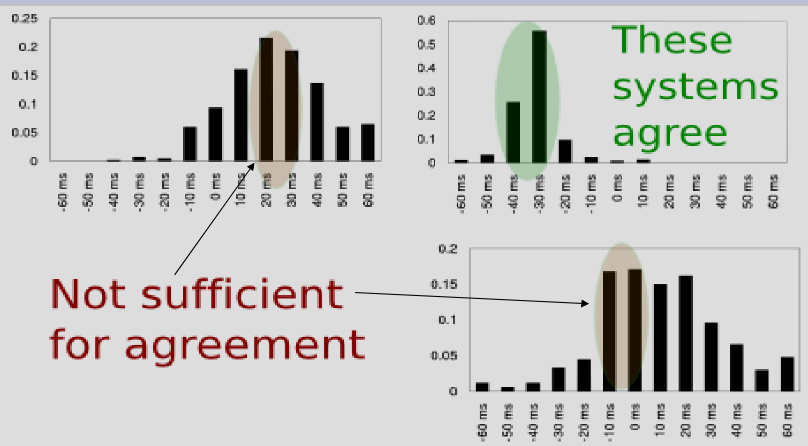
The location of cuts don't matter (much) as long as they are consistent.

How to measure?



Repeat for all pairs; count the number of agreements.

Algorithm



...then count the number of systems that agree for that particular transition.

Training Data and Technologies

Training Data:
- Ad hoc, simple read sentences and citation form words.
- 23,000 utterances, 48,000 total words, vocabulary size=16,000.
- 16kHz sampling rate.

Technology:
- HTK (Cambridge UK) toolkit, monophone models.
- “short pause” phoneme inserted between words.

Three groups of systems:
- * MFCC front end, 20ms window
- * LP-Cepstra front end, 24ms window.
- * Auditory description vectors front end
  - Cube-rooted erb-wide filters
  - Edge detectors
  - Voicing detector, etc
  - Reduced via principle components analysis from 51 dimensions to 19

48 alignment systems in total.

HMM details:
- 1, 2, 4, or 8 Gaussians per state
- 2 or 3 states per phone (strict left-to-right)
- 4 or 5 states per phone (allowing skips).

Results

Classes of phonemes:

Class	Abbr.	British English Phonemes (SAMPA)
Nasal	Nas	m, n, N
Plosive	Plo	b, d, g, k, p, t
Affricate	Aff	dZ, tS
Fricative	Frc	D, S, T, Z, f, h, s, v, z
Vowel	Vow	@, A, E, I, O, U, V, {, 3, i, u, Q
Approximant	App	r, j, l, w
Diphthong	Dip	I@, U@, aI, aU, E@, eI, OI, @U
Silence	Sil	silence, short inter-word pause

Compute the histograms for all transitions from one class to another.

Why classes? Some phoneme-phoneme transitions are quite rare.

Some transitions have vastly more agreement than others.

Number of agreements out of 1128 possible pairs.

Transitions Class A → Class B		Class B							
		Plo	Aff	Frc	Nas	App	Vow	Dip	Sil
Class A	Plo	19	39	70	84	166	411	469	57
	Aff	369	–	25	–	63	688	753	123
	Frc	438	86	22	456	133	423	537	116
	Nas	241	632	247	5	52	332	396	84
	App	208	156	251	60	5	26	9	103
	Vow	332	365	254	182	8	1	1	78
	Dip	361	437	260	170	6	0	4	77
	Sil	923	395	299	817	713	513	516	–

Results

Class	Abbr.	British English Phonemes (SAMPA)
Nasal	Nas	m, n, N
Plosive	Plo	b, d, g, k, p, t
Affricate	Aff	dZ, tS
Fricative	Frc	D, S, T, Z, f, h, s, v, z
Vowel	Vow	@, A, E, I, O, U, V, {, 3, i, u, Q
Approximant	App	r, j, l, w
Diphthong	Dip	I@, U@, aI, aU, E@, eI, OI, @U
Silence	Sil	silence, short inter-word pause

Transitions Class A → Class B		Class B							
		Plo	Aff	Frc	Nas	App	Vow	Dip	Sil
Class A	Plo	19	39	70	84	166	411	469	57
	Aff	369	–	25	–	63	688	753	123
	Frc	438	86	22	456	133	423	537	116
	Nas	241	632	247	5	52	332	396	84
	App	208	156	251	60	5	26	9	103
	Vow	332	365	254	182	8	1	1	78
	Dip	361	437	260	170	6	0	4	77
	Sil	923	395	299	817	713	513	516	–

* Transitions between similar phonemes are very uncertain

Results

Class	Abbr.	British English Phonemes (SAMPA)
Nasal	Nas	m, n, N
Plosive	Plo	b, d, g, k, p, t
Affricate	Aff	dZ, tS
Fricative	Frc	D, S, T, Z, f, h, s, v, z
Vowel	Vow	@, A, E, I, O, U, V, {, 3, i, u, Q
Approximant	App	r, j, l, w
Diphthong	Dip	I@, U@, aI, aU, E@, eI, OI, @U
Silence	Sil	silence, short inter-word pause

Transitions Class A → Class B		Class B							
		Plo	Aff	Frc	Nas	App	Vow	Dip	Sil
Class A	Plo	19	39	70	84	166	411	469	37
	Aff	369	–	25	–	63	688	753	123
	Frc	438	86	22	456	133	423	537	116
	Nas	241	632	247	5	52	332	396	84
	App	208	156	251	60	5	26	9	103
	Vow	332	365	254	182	8	1	1	78
	Dip	361	437	260	170	6	0	4	77
	Sil	923	395	299	817	713	513	516	–

* The table is somewhat symmetrical (ie A|B agrees if B|A agrees) but with some dramatic exceptions: pairs involving silence and plosives.

Why silence? Sentences start sharply and end gradually.
Why plosives? The beginning of a plosive is much different from the end.

Inside the classes: a look at individual phonemes

The fewest agreements (<5% agree):

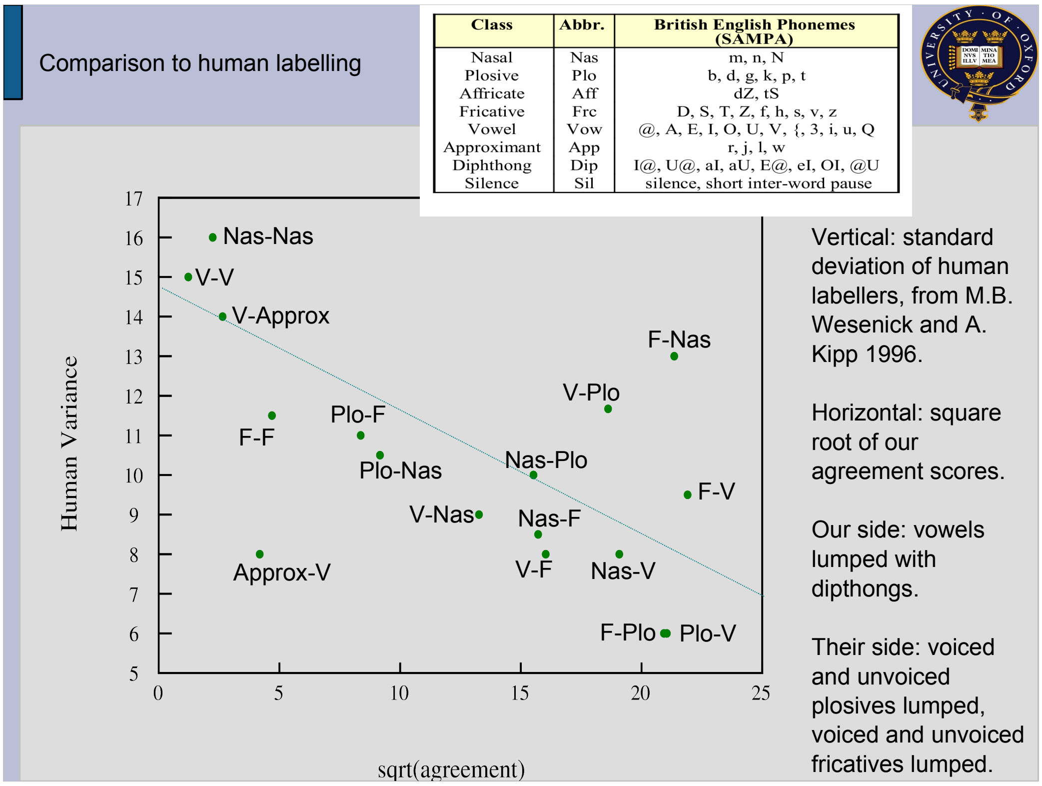
- /k/ or /v/ → Silence
- /aI/ or /aU/ → /@/
- …

Most agreements (>99% agree):

- Silence → /S/, /d/, /b/, …
- /s/ → /@U/, /eI/, /aI/, /O/, …
- ...

Transitions Class A → Class B		Class B							
		Plo	Aff	Frc	Nas	App	Vow	Dip	Sil
Class A	Plo	19	39	70	84	166	411	469	57
	Aff	369	–	25	–	63	688	753	123
	Frc	438	86	22	456	133	423	537	116
	Nas	241	632	247	5	52	332	396	84
	App	208	156	251	60	5	26	9	103
	Vow	332	365	254	182	8	1	1	78
	Dip	361	437	260	170	6	0	4	77
	Sil	923	395	299	817	713	513	516	–

Individual phoneme results are generally consistent with class-based results.



Comparison to human labelling
Class | Abbr. | British English Phonemes (SAMPA)
Nasal | Nas | m, n, N
Plosive | Plo | b, d, g, k, p, t
Affricate | Aff | dZ, tS
Fricative | Frc | D, S, T, Z, f, h, s, v, z
Vowel | Vow | @, A, E, I, O, U, V, {, 3, i, u, Q
Approximant | App | r, j, l, w
Diphthong | Dip | I@, U@, aI, aU, E@, eI, OI, @U
Silence | Sil | silence, short inter-word pause
Human Variance
sqrt(agreement)
Nas-Nas
V-V
V-Approx
F-Nas
V-Plo
Plo-F
F-F
Plo-Nas
Nas-Plo
F-V
V-Nas
Nas-F
Approx-V
V-F
Nas-V
F-Plo
Plo-V
Vertical: standard deviation of human labellers, from M.B. Wesenick and A. Kipp 1996.
Horizontal: square root of our agreement scores.
Our side: vowels lumped with dipthongs.
Their side: voiced and unvoiced plosives lumped, voiced and unvoiced fricatives lumped.

Conclusions

- Substantial differences between different boundaries
- Some correlation with human errors.
- Precision tends to be lower for similar phonemes
- Precision best at beginnings,
- Worst at ends.
- When evaluating alignment systems, weight the boundaries.